UNIVERSITY OF LIVERPOOL

Towards cancer diagnosis via tissue discrimination using various infrared spectroscopy techniques.

by

James Ingham

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
School of Physical Sciences
Department of Physics

June 19, 2018

"Blood, Sweat and code."

Steve Barrett

UNIVERSITY OF LIVERPOOL

Abstract

School of Physical Sciences

Department of Physics

Doctor of Philosophy

by James Ingham

One of the largest challenges within modern medicine is the increase in global cancer rates especially in western countries, which is often attributed to ageing populations, dietary and lifestyle changes. One of the fastest growing cancers within the western world is oesophageal cancer, which without reliable early diagnosis is often fatal due to the cancer spreading. It is therefore crucial for the development of reliable methods and tools for early cancer detection. This is especially important for those who are most at risk, which includes Barrett's oesophagus patients.

Infrared (IR) spectroscopy techniques have been proven capable of gaining large amounts of information on the chemical composition of biological samples. This thesis therefore focuses on using a variety of IR spectroscopy techniques, including Fourier transform infrared spectroscopy (FTIR) and scanning near-field optical microscopy (SNOM) to image oesophageal samples, tissue biopsies and cell line samples. The thesis demonstrates how machine learning algorithms can be used in conjunction with FTIR to provide a quick, non-baised tissue diagnostic method, free from the issues associated with current histology techniques. As well as focusing on the processing of FTIR data, the thesis will assess the ability of an aperture SNOM to image biological samples as it is able to achieve diffraction limit breaking spatial resolutions and has to potential to give previously impossible insights.

Acknowledgements

I would like to first thank David Martin and Steve Barrett for the constant support throughout my entire PhD. I would also like to thank all the members of the Liverpool SCAnCan group, including Peter Weightman, Michele Siggel-King, Caroline Smith, Paul Harrison, Paul Unsworth and Timothy Craig, who have all been great at guiding and encouraging me throughout the last four years. A special thanks to Caroline who persevered through reading my many spelling mistakes when helping me through my corrections. Thank you to Andrea Varro and Mark Pritchard who supplied the samples used throughout this thesis. I want to thank Peter Gardner, Mike Pilling and Paul Bassan who form the Manchester contingent of the SCAnCan group and managed to squeeze me into the often full FTIR schedule. I must thank the ALICE team based at Daresbury, especially including Neil Thompson, Yuri Savaliev, David Dunning, Ben Shepherd and many others who gave up their weekends to operate ALICE and those who acted as second commissioners. A special thanks to Mark Surman who helped make the SNOM experiments possible.

Contents

List of Figures

List of Tables

Abbreviations

AFM	Atomic Force Microscopy
ALICE	Accelerators and Lasers In Combined Experiments
ATR	Attenuated Total Reflectance
CA	Cluster Analysis
EM	Electromagnetic
FEL	Free Electron Laser
FPA	Focal Plane Array
FTIR	Fourier Transform Infrared
FTS	Fourier Transform Spectroscopy
FWHM	Full Width at Half Maximum
H&E	Haematoxylin and Eosin
IR	Infrared
MA	Metric Analysis
MCT	Mercury-Cadmium-Telluride
PCA	Principal Component Analysis
PDF	Probabiltiy Density Function
QCL	Quantum Cascade Laser
SEM	Scanning Electron Microscope
SNOM	Scanning Near-Field Optical Microscopy
SPM	Scanning Probe Microscopy
STM	Scanning Tunnelling Microscope
TEM	Transmission Electron Microscope
TERS	Tip Enhanced Raman Spectroscopy

Chapter 1

Introduction

1.1 Cancer and histology

Cancer is one of the major causes of death throughout all demographics around the world, responsible for $\approx$ 16% of all deaths globally [1]. As the World Health Organisation is predicting that incidence rates will increase by over 50% by 2020 [2] it is only going to become a greater issue in the future. This dramatic increase is attributed to many factors including ageing populations and detrimental dietary and lifestyle changes [3]. Cancer research is therefore one of the major areas of active research involving many scientific disciplines, often working within collaborations. One such collaboration was the SpectroChemical Analysis for Cancer (SCAnCan) group funded by EPSRC, which consisted of many collaborators from various scientific backgrounds across multiple UK universities. The author of this thesis was a member of SCAnCan and the work shown within this thesis was carried out under the SCAnCan research remit. The aim of most cancer research is to advance the effectiveness of treatments for various cancers, but areas such as diagnostics, early screening and development of the understanding of key processes within cancer are also funded. Cancer Research UK for example spent over £430 million during 2016/2017 on research [4]. Although much work has already been done with advances resulting in survival rates greatly increasing in

the last few decades for some cancer types, other forms of cancer have shown little improvement, which highlights the broadness and often oversimplification of the term 'cancer research'. The work within this thesis aims to focus on the problems involved in diagnostics and early screening of cancer to ultimately help deliver effective and reliable methods for clinicians to best diagnose their patients. This is a critical process as it has been shown in many cancers that effective early diagnosis which leads to treatment can dramatically increase the chance of survival. For example, a patient diagnosed with early stage colon cancer has a 90% 5 year survival rate compared to 11% if diagnosed at a later stage of cancer [5], similar figures are shown for breast cancer with 90% and 15% 5 year survival rates quoted for early and late stage cancer respectively [6].

Cancer has been known to exist since as early as 1600 BC as the ancient Egyptians documented cases of bone and breast cancer reporting them as being untreatable [7]. There are even records which demonstrate that skin cancer tumours were removed in much a similar fashion as is done today in modern medicine. The word cancer was first used by physician Hippocrates and is derived from the Greek word 'carcinos' meaning crab. It was named so because he believed the carcinoma tumours he was studying at the time looked similar to a crab as they often had a central mass with thinner extruding structures. There have been many ideas proposed throughout history as to the causes of cancer including the humoral theory, which was common in the middle ages [7]. It stated that there are four main fluids within the body and any imbalance of these fluids would cause a variety of ailments including cancer. Up until the late 18^{th} century doctors even suggested that cancer was in itself contagious and a person having cancer of any kind could be spread via parasites. Some forms of cancer can be caused by parasites but the cancer can not be spread from one person to another.

Today it is understood that cancer occurs when normally functioning cells have their DNA damaged and so begin to behave abnormally, usually resulting in an increased rate of cell division which causes them to grow out of control. For cancers which occur in tissue the cancerous cells often form a solid mass of abnormal cells, described as a tumour, and are even capable of forming their own blood

supplies to fuel their growth [8]. Tumours come under two main categories benign and malignant, with the former being tumours that have stopped growing and are no longer life threatening and the latter being cancerous masses which are likely to continue growing and to spread to other parts of the body (metastasis). Healthy cells can often have their DNA damaged, but in most cases it is quickly repaired by internal mechanisms within the cell resulting in no further complications. There are even processes in place to force damaged cells to die (apoptosis) so they don't spiral out of control, which can be signalled within the cell and also by surrounding cells. When a cell becomes cancerous this regulation of the cell can be ignored, resulting in it proliferating rather than undergoing apoptosis. Cancer can be caused by a large variety of reasons including genetics, but it is most often caused by environmental factors [9]. Through the detailed study of many cancer patients, genes called 'oncogenes' have been found which when damaged are prone to resulting in cancer. It is important to clarify that 'cancer' is an umbrella term for over 200 diseases, with each having its own challenges, diagnostics and treatments, with some occurring more commonly than others and they often have large discrepancies in the survival rates [10].

The first stage of any cancer treatment is for the clinician to get a clear indication as to the condition of the cancer, how severe it is, what type of cancer is present, the aggressiveness and also if it has spread to other areas of the body. There are many cutting edge instruments available in modern medicine such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans and Positron Emission Tomography (PET) scans which allow for the imaging of soft tissue within a patient. But the current standard for the evaluation of potentially cancerous tissue is still pathology, which has been an established and largely unchanged as a technique for hundreds of years [11]. Within the National Health Service (NHS) over 70% of all disease diagnostics use some form of pathology, equating to over 800 million tests taking place annually within the UK [12]. Pathology is the study of biological samples associated with various diseases and is often done using optical microscopy with the aim of ascertaining key information on the sample. Histopathology is a branch of pathology focused on the study of

tissues specifically to diagnosis disease. Histopathology is often done *ex vivo* with biopsies being taken by endoscopes for detailed study under laboratory conditions. It is important for histologists to determine the stage and aggressiveness of any cancer present in the biopsy as it can often show different characteristics from patient to patient. Some well-studied cancers have a specific grading system in place to accurately describe the nature the cancer at various stages of development, such as the Gleason scale for prostate cancer [13]. The grading is done by comparing the biopsies to community outlined standards for the general appearance of previously graded tissues.

Optical microscopes which operate within the visible portion of the electromagnetic (EM) spectrum are used all over the world and have been a key fundamental within science for hundreds of years. There are many advantages of using these microscopes as they are relatively cheap, simple to use and require a light source which is easy to produce. To properly view the structures within the biopsies with optical microscopes they have to undergo a standardised procedure to prepare them for study. This process has many steps and starts with the tissue biopsy being chemically fixed, usually with formaldehyde, which stops the sample from changing chemically and decaying. The next stage is to embed the biopsy in a paraffin wax to support and preserve the sample. An alternative to embedding the sample is wax is to cryofreeze it, but this is often not ideal as it can result in morphological degradation [14]. The next step is to slice the biopsy into very thin sections appropriate for viewing in a microscope as they need to be transparent. This is done using a microtome, which can cut the sample into thin wafers of between 0.5-100 µm thick, for the studies within this thesis the tissue samples were 5 µm thick. The slices are then mounted onto standardised transparent slides which allow the sample to be easily moved and placed in the microscopes. By using microscopes to study the interaction of the light with the tissue slices, detailed information about the samples morphology, the structures within the tissue, can be gathered.

This is not generally adequate for most histopathology diagnostics as tissue samples tend to show very little contrast. Because of this dyes have been developed

to stain the tissue samples various colours depending on the molecules present, thereby adding additional contrast to the microscope images, as demonstrated in Figure 1.1. The most commonly used stain in histology is Hematoxylin and Esoin, commonly referred to as H&E, which highlights areas of the tissue dense in nucleic acid blue and areas rich in protein pink. The addition of this contrast allows for much greater tissue differentiation as the microscope image now indicates the distribution of some of the molecular constituents within the sample. This practise is referred to as the current 'gold standard' and gives a pathologist enough information from which they can often make an informed prediction on the state of the tissue.

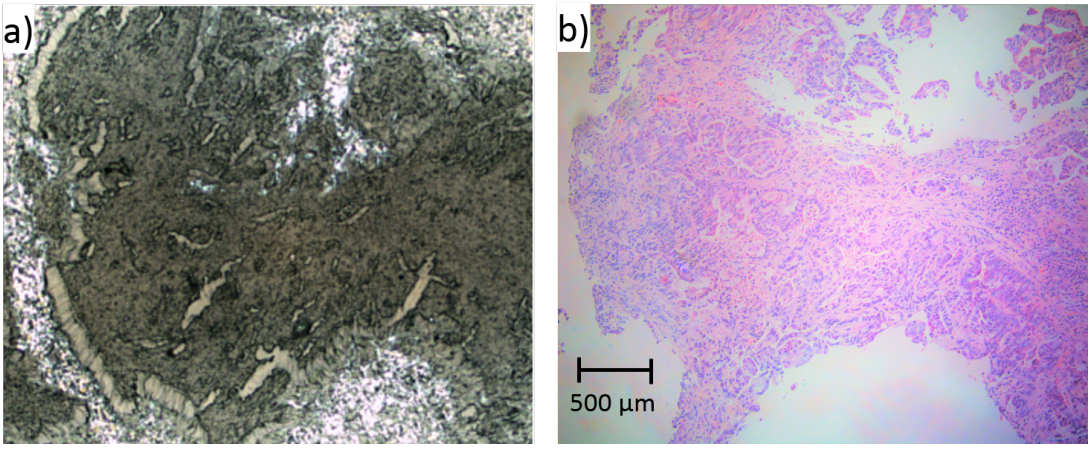

FIGURE 1.1: An image comparing two different slices from a Barrett's oesophagus tissue biopsy; a) is an image of an unstained slice which is still embedded within wax, and b) an adjacent slice from the same biopsy which has been dewaxed and stained with H&E.

Even though modern histopathology is a powerful tool capable of providing important information for disease diagnostics, it isn't a perfect technique and therefore a drive for different approaches is present. One issue is that it can often take a large amount of time, $\approx$ 72 hrs at least depending on its seriousness [15], from the biopsy being taken to the pathologist having classified it and informing the clinician who can then determine the optimal treatment. Because the process is somewhat laborious with the many stages needed to prepare the sample, it is not possible to carryout the diagnostics during an operation which would be ideal. Another issue is that a single slice tends to be only stained by one dye so there is

a limited amount of chemical information available to be studied. Though this is somewhat remedied by staining multiple slices from the same biopsy, which while not being exactly the same will share common structures if taken close to each other, it is still a limitation.

The major inherent weakness though is that it is fundamentally a subjective classification method. Although stains allow for more contrast and microscopes allows for greater detail the diagnosis is still based on the pathologists interpretation of the structures and appearance of the stained areas within tissue samples. As the grading of a sample is based on the pathologist comparing it to the community outlined criteria it is clear how multiple pathologists could disagree on the final grading, which has been documented to regularly occur [16–18], with some studies showing pathologist only agreeing 30-50% of the time for low grade cancers. An incorrect classification could potentially lead to either very risky unnecessary surgery or a cancerous tumour being left to spread to the point that it is no longer treatable. It is these flaws which limit current histopathology as it doesn't fully meet the needs of clinicians who need a quick and reliable classification methodology.

One promising alternative approach is chemical imaging, a large and still developing field, which looks to be capable of beneficially complementing current histopathological practices. Chemical imaging techniques are able to capture the chemical content of the samples without the need for dyes and can gain information on a larger number of molecular species than previously possible. A considerable branch of chemical imaging is Infrared (IR) spectroscopy which has become a well developed field and is a forerunner as a viable option for the future of disease diagnosis, with many examples of successful studies [19–26]. These examples were capable achieving of both high sensitivity and specificity at distinguishing healthy and cancerous tissue for a range of cancer types. Techniques such as Fourier Transform InfraRed (FTIR) spectroscopy and Raman spectroscopy are the forerunners within chemical imaging as they have been refined in understanding and instrumentation and are most commonly used.

IR spectroscopy still relies on the interaction of light with the sample similar to the standard visible light microscopy, but instead uses an IR light source to probe the sample. Techniques such as FTIR benefit from the strong absorption of IR light by biological materials as they contain organic molecules such as proteins, nucleic acids, lipids and carbohydrates, which have molecular vibrational states with similar energy gaps as the energy of the IR photons. Absorption features within IR spectra occur because the photons are absorbed by the molecules causing particular bonds within the molecule to vibrate. As each species of molecule will tend to have specific bond types it will also have a unique set vibrational states and so the IR absorption spectra will have distinctive features, often called IR fingerprints. For FTIR experiments on biological samples a mid-IR light source with wavelengths ranging from 2-14 µm is used as they contain the most important features [27].

IR spectroscopy is therefore able to gain detailed information into the chemical constituents without using labels and an indication to the relative abundance of the many molecular species within the sample, as the IR spectra is an amalgamation of all the component fingerprints. This results in IR spectroscopy techniques potentially outperforming the current dyes used in histology as it can simultaneously detect the presence of a greater number of molecular species. Although spectroscopy is primarily focused on the interaction of light and the sample at various wavenumbers, by using a two dimensional Focal Plane Array (FPA) detector with an instrument such as an FTIR, multiple spectra from different spatial positions can be collected simultaneously and hence is a combination of both spectroscopy and microscopy.

A single FTIR image can therefore contain many thousands of spectra, with each spectrum containing information at over a thousand different wavenumbers. Therefore the advantage of IR spectroscopy for cancer diagnosis is the much greater insight into the sample composition, which should allow for better diagnostics and also potentially an insight into the subtle chemical differences between the various stages of cancer.

FTIR has been already been applied to the study of many types of cancer including breast [28, 29], prostate [30], oral [31], RNA taken from brain tumours [32], colon [33], cervical [34, 35]. Mordechai *et al* studied FTIR datasets of both cervical cancer and melanoma against samples of nonmalignant images, with the aim to discover potential biomarkers which distinguish between healthy and cancerous tissues [36]. They studied the IR spectra with a focus on the features relating to important biological molecules such as RNA, DNA, phosphates and glycogen. For cervical cancer they found the relative amounts of glycogen to be a potential biomarker between cancerous and healthy tissue. This wasn't the case for melanoma which instead indicated that the relative amounts of RNA and DNA was a distinguishing factor. Although they only had a limited amount of data within this work and so is not currently robust enough for medical applications yet, it demonstrates the potential strength of FTIR spectroscopy as a powerful tool for tissue discrimination.

A study by Gazi *et al* [37] combined the study of FTIR spectroscopy with traditional histopathology to grade various spectra on the Gleason scale, which ranges from 2-10 with a higher grade indicating a more severe case. FTIR spectra were associated to particular Gleason grade, which was done by assessing the equivalent area on a associated H&E stained slice from the same biopsy, and used to construct a diagnostic classifier capable of grading the IR spectra of prostate tissue. The classifier was then tested on spectra which were put in to three 'bands' depending on the associated Gleason grade; those with a grade lower than 7, those with a grade of 7 and those with a grade higher than 7. The classifier would then predict the grade of each of the spectra and then could compare this grade against the associated grade from the H&E slice. For spectra with an associated grade of less than 7 they achieved a sensitivity and specificity of 70% and 89% respectively, which was similar to the sensitivity and specificity of grade higher than 7 which were 71% and 89%. Finally for tissue associated to a Gleason grade of 7 the classifier achieved a sensitivity and specificity of 78% and 81%. Although the performance of this classifier is similar to that current histopathology achieving around 70%, it is a clear indicator that the internal chemistry with which the

FTIR is sensitive to is a viable metric capable of grading tissue to a predefined scale, which was not developed with IR chemical imaging in mind.

One of the previous drawbacks of IR imaging was the limited field of view. This was improved by the implementation of FPAs to record multi-spectrum images, but this is still often limited to an area of $\approx 700\,\mu m\ x\ 700\,\mu m$. As biopsies can be considerably larger than this it was seen as a major limitation of FITR. A study by Bassan *et al* [38] showed that by using a mosaic method of combining multiple scans together in an automated manner, a slice from an entire prostate which was 4 cm x 5 cm in size could be studied. The resultant image consisted of 4047 individual FTIR image stitched together to produce a single large dataset containing 66 million pixels and took around 14 hours to complete. Although 14 hours is a large amount of time to image a single sample it does highlight that FTIR instruments are not limited to small fields of view and have potential for imaging whole tissues. In the future it is foreseeable that techniques such as FTIR will be combined with a prescan method which can quickly assess the sample on a large scale and highlight areas of interest which are then to be studied in detail by the FTIR instrument, therefore minimising the scan time.

As demonstrated FTIR has already established itself as powerful tool capable of gaining important information to characterise biological samples, the major challenge is therefore in how to best process and analyse the data to meet the needs of the clinicians and researchers alike. As FTIR datasets are very large with a single image often containing over 24 million data points and most studies will usually use multiple images, it is too much information for a person to reasonably handle. Therefore computational techniques are needed to both process and interpret the data, which has recently become possible over the last decade as computational power has grown considerably. Machine learning (ML) is a field which focuses on using statistical methods to allow computers to 'learn' and classify data. ML is one of the fastest growing fields having been applied to all manner of tasks including self driving vehicles [39, 40], finding trends and predicting future changes in the financial market [41, 42] and object recognition needed for artificial intelligence computer vision [43].

ML is has also become a considerable tool within medicine as an ever increasing number of imaging tests are producing large digital images which benefit from the strengths of ML. A key usage of ML is as a classifier capable of labelling input data depending on distinguishing features. There are many types of classifiers available each with their own potential strengths and weakness. Most come under two categories of being either supervised and unsupervised. The key difference between them being that supervised learning uses previously labelled example data to find reliable distinguishing features between sample types (eg. tissue type A and tissue type B), while unsupervised learning methods try to infer the structural differences within the datasets with out prior labelling/learning.

ML doesn't just help with handling large amounts of data it is also capable of finding very small trends present within the IR spectra, which is a nontrivial problem as most IR spectra of biological samples often appear very similar since they tend to contain the same molecules. The area of automated data classification is a key part of the rapidly expanding field of ML, already having been applied to many biological sample classifications [44–48].

Lasch *et al* [49], studied FTIR images of colorectal adenocarcinoma samples using three different methods of cluster analysis which are unsupervised methods, k means, fuzzy C-means and hierarchical clustering. Cluster analysis in general aims to group spectra together based on how similar the features within the IR spectra appear. This means that spectra within a given cluster, also called a class, appear to be more similar than spectra belonging to another class, meaning that the inter-class variation is larger than the intra-class variation within the cluster. It does this by mapping the data in n dimensions, where n is the number of wavenumbers within the study. Spectra which have similar profiles will therefore be plotted closely in the n dimensional map and hence clustered together, Figure 1.2 shows a simple example with 3 clusters shown in 2 dimensions, with each cluster coloured differently. Once each spectra has been labelled with a given class a false colour image which shows the assigned label of each of the spectra spatially can be made which allows for the tissue structure to be studied. Within the studied it was found that all three methods worked well at discriminating the different tissue

types with hierarchical clustering performing the best, but also taking the longest to process at ≈ 4.5 hrs. They concluded that all the methods used increased the information content of the IR images giving a greater insight into the structures within the sample.

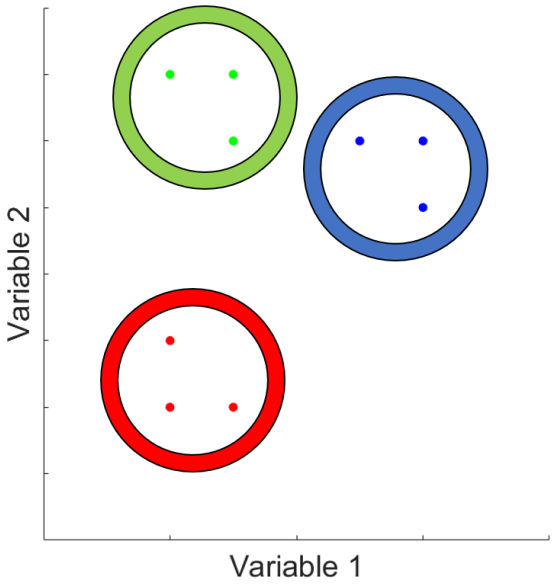

FIGURE 1.2: A simple example to show the premise of cluster analysis. The data points shown above have been clustered into 3 classes and are distinguished by their colour. Data points within the same class are deemed as being more similar than the spectra not within the class.

In a study by Zawlik *et al* [50], FTIR images of triple negative breast cancer (TNBC), which is the most aggressive example of breast cancer, were studied with the use of principle component analysis (PCA) which is another unsupervised method. TNBC is a subtype of epithelial breast tumour and TNBC patients often respond poorly to the standard chemotherapy treatment. Zawlik *et al* therefore took FTIR images of tissue biopsies taken before and after chemotherapy. By using PCA, which is a method that aims to find the major differences between the spectra, they determined that discriminatory features were present within the spectra. These differences were consistent with the pathologic and clinical responses to chemotherapy and hence indicated the potential for IR imaging systems to be used in testing the effectiveness of chemotherapy treatments. PCA works by converting the inputted data into several principle components which

are essentially the key underlying structures within the IR spectra. They are ordered in such away that the first principle component is the one where there is the most variation, the second having the second most variation etc. PCA is therefore capable of finding both differences between spectra but also compresses the data into key information needed for spectra classification. Because of this compression PCA is often used before other ML techniques to reduce the large data sets to more reasonable sizes, while still keeping the critical information.

An example of PCA being combined with another classification method was shown by Kaznowska *et al* [51], who used PCA with linear discriminant analysis (LDA) to classify various examples of colon tissue. Within the study they compared spectra generated by FTIR of healthy, cancerous and post-chemotherapy colon samples. Within the study they were able to achieve good separation between the healthy and pre/post chemotherapy samples at wavenumbers associated to key biological molecules, with a particularly strong discrimination at 1385 cm^{-1} which they concluded may become a key biomarker in the future. This study again confirms that FTIR spectroscopy is a reliable source of important biometric information which relates to the internal chemical composition of tissue in various stages. Similar to the Zawlik *et al* study, the results imply that techniques such as FTIR may be a useful tool to facilitate the monitoring of the efficiency of treatments such as chemotherapy for cancer patients.

It is not just unsupervised ML techniques which have been applied to IR spectra as Pilling *et al* [24] showed that a random forest (RF) classifier, a supervised method, can be applied to sample discrimination as they used it on FTIR data of prostate samples to achieve extraordinary discrimination. RF creates multiple decision trees which aim to categorise the IR spectra based on their predefined labels. Once a model has been generated capable of discriminating the previously labelled example spectra, it can be applied to spectra which it has not previously processed and make a prediction as to the class of the spectra based on its spectral features. In this study Pilling *et al*, achieved discrimination of normal epithelium, malignant epithelium, normal stroma and cancer associated stroma at over 95%. Along with the excellent discrimination they also stained the samples with H&E

before imaging them and also proved that the staining didn't contribute to the chemical discrimination. This is important as it shows that FTIR as a technique can easily work side by side with other currently used histopathology techniques. The results of this study show the strengths of a supervised method which allow the ML algorithm to 'learn' from previously labelled samples and find the key distinguishing factors which can then be used as important biomarkers for labelling images in the future.

Fabien *et al* [52] applied another type of supervised learning method, artificial neutral network (ANN) analysis to breast cancer samples. With this study they collected FTIR data of four types of tissue; fibroadenoma, ductal carcinoma, connective tissue and adipose tissue for both benign and malignant lesions in breast cancer. They were able to achieve a discrimination success of 90%+ between the benign and malignant samples. This indicated the strength of IR imaging as it allows for a reliable and rapid diagnosing of tissue samples which is the fundamental aim for any histopathological tools. This study also demonstrated that these techniques are not just capable of detecting the difference between healthy and cancerous tissue but also between benign and malignant tumours, which would clearly be a very useful tool for ascertaining the seriousness of a tumour quickly.

The development of an algorithm capable of processing large numbers of IR spectra while also finding key hidden biomarkers is a considerable part of this thesis and will be evaluated using FTIR data of both tissue and cell line samples. The aim was to produce an algorithm which is capable of meeting both the needs of a clinician and also researchers. Clinicians/histopathologists desire a diagnostic tool which is capable of rapidly labelling the contents of a tissue sample, flagging the presences of any diseased tissue and if there is, to potentially grade/evaluate the severity. This ML method would only need to be a 'blackbox' approach as the final answer is more important than the details of how the answer was produced, if it has already been proven to be reliable. On the other hand researchers need a tool capable of providing a greater insight into the chemical composition of the diseased tissue and an algorithm which provides clear information on the structural and chemical differences between any samples provided. This knowledge could enable

a greater understanding of the fundamental processes at work within cancer and may ultimately lead to better prevention, diagnostics and treatments.

As with many rapidly expanding areas of dynamic research such as ML coupled with chemical imaging, there are problems that need to addressed before they are adopted fully into medical standard practises [53, 54]. These are not so much fundamental problems but rather community issues, such as the lack of standardised sample preparation routines, accepted outlines for experimental practices and no general agreement as to the best ML methods to use. This wide range of ideas may help as many groups are tackling the problem of tissue diagnostic via IR spectroscopy in so many different directions which may ultimately lead to improvements in procedures, but this lack of consensus within the community is also slowing the rate at which these same techniques are able to push towards general adoption. There is also the issue that presently there are very few IR spectroscopy studies that come close to the size of standard clinical trails and therefore gaining the acceptance of the medical community at large would currently be very hard.

Optical instruments which use conventional optical lenses are fundamentally limited in their spatial resolution. This limit is imposed by the diffraction limit and affects all far-field optical instruments such as standard microscopes as well as chemical imaging instruments such as FTIR and Raman. The diffraction limit is explained in detail in Chapter 2, but can be generalised to Equation 1.1, where d is the best possible resolution and λ is the wavelength of the probing light.

$$d = \frac{\lambda}{2} \tag{1.1}$$

For mid-IR experiments such as those shown within this thesis this limits the spatial resolution to $\approx$ 2.5-4 µm, which is somewhat worse than the optimal spatial resolution of microscopes using visible limit which have a limit of around 0.2-0.35 µm. Therefore when moving from the visible portion of the EM spectrum to the IR region, when using far-field techniques spatial resolution is sacrificed. It is important to note that these values are also ideal values and that in reality it can

be very hard to approach these values due to defects in optics and misalignments. These lower resolutions for IR imaging may still be adequate for some situations but not if spatial resolution is critical, which is common for many surface science experiments, the study of thin membranes and imaging of very small intercellular structures. This has led to the development of instruments such as Scanning Near-field Optical Microscopy (SNOM) and Tip Enhanced Raman Spectroscopy (TERS), which have the potential of IR imaging at much greater resolutions as they exploit the near-field interaction [55]. Pohl *et al* demonstrated the first example of diffraction 'breaking' resolution using a visible light source and achieved a resolution of 25 nm which equates to $\approx \frac{\lambda}{20}$ [56].

Richter *et al* [57] used TERS to image the proteins and lipids on the surface of a cancerous colon cell (HT29). TERS is able to beat the diffraction limit by localising the electromagnetic field at the apex of a very sharp metallic tip which has a surface-enhanced Raman spectroscopy (SERS) coating. By using a near-field technique such as TERS they were able to achieve a spatial resolution of between 10-20 nm. By using a form of data analysis they were able to separate the contributions from both the lipids and proteins within the IR spectra. This allows for the independent mapping of both molecular species at a nanometre scale which is fundamentally impossible with far-field techniques such as FTIR, Raman and standard optical microscopy.

In another TERS study by Wood *et al* [58] they were able to demonstrate that TERS is capable of mapping the presence of hemozoin crystals, a product produced by some blood feeding parasites, within a sectioned malaria infected red blood cell. They were able to study the features within the IR spectra produced by TERS and compare these to the expected features of the hemozoin crystals and were therefore capable of determining the presence of such crystals within the cell. Again they were able of achieving incredibly high spatial resolution while still capturing the desired IR spectra which allow for the mapping of various chemicals. The authors comments that they foresee approaches such as TERS to lead the way to a novel drug screening modality for the detection of drugs binding

to the hemozoins surface within cells without the hemozoin crystals needing to be extracted first.

TERS is not the only high spatial resolution IR imaging technique to gain traction as SNOM is also seeing a surge in attention. Cricenti *et al* [59], demonstrated that an aperture SNOM instrument coupled with a IR free electron laser (FEL) is capable of imaging thin film samples at a spatial resolution of 200 nm which is $\approx \frac{\lambda}{35}$. They first used a FTIR to gather general spectra for both the sulphur and nitrogen oxide compounds. By using these IR spectra a wavenumber could be chosen which is shown to behave differently for both compounds, which is this example was 1438 cm^{-1}. This demonstrates the practise of using the wide spectral information available in far-field techniques to inform the experiments of near-field studies, this also took place within the SCAnCan experiments. By comparing the image produced while using 1438 cm^{-1} against the image when imaging with light at a wavelength with poor light-sample interaction such as 1515 cm^{-1}, chemical specificity is demonstrated as there is clear contrast in the image produced when using 1438 cm^{-1}, showing the chemical variation on the surface and very little when using 1515 cm^{-1}. This highlights again as with TERS the powerful potential of near-field techniques to gain chemical sensitive images of samples at very high resolutions.

A recent study by Halliwell *et al* [60], who was also member of the SCAnCan group, used the same aperture SNOM which is described in detail later within the following chapter to image several types of cervical samples. The samples were images at four wavenumbers associated to key biological cell components; DNA (1225 cm^{-1}), Amide II (1550 cm^{-1}), Amide I (1650 cm^{-1}) and lipids (1750 cm^{-1}). Within the study they were able to demonstrate that by using ML techniques already discussed such as PCA and PCA-LDA it is possible to achieve a high level of discrimination between the samples. This ability to discriminate between various tissue samples is akin to the far-field techniques described earlier but with instruments such as SNOM it is possible to accomplish the same feat while also maintaining a very high resolution.

As SNOM instruments are not limited to the optical limitations such as the diffraction limit any light source can in principal be used as long as the apparatus involved in the instrumentation are capable of effectively transmitting the light. This means that SNOMs have not only been assigned to IR wavelength, highlighted by studies such as Perner *et al* [61] who used visible light to image the effects of varying the oestrogen concentration of breast cancer cells. Although the SNOM did not have access to the wealth of information available in IR imaging it was still able to gain high resolution images of the morphology of the sample which is what was desired within this experiment.

In a study by Kaupp *et al* [62], a variant of SNOM has demonstrated which uses an apertureless tip unlike the previous examples which used an aperture SNOM fibre, the difference will be discussed in detail in the following chapter. In essence the difference is in how the SNOM tip is sensitive the light-sample interaction. Within this study they were able to again achieve a high resolution using a SNOM while also demonstrating clear differences between the images of healthy and cancerous bladder samples.

Instruments such as TERS and SNOM are still in development and are still in the process of solidifying their position within the imaging field. As they have been shown in several cases to beat the diffraction limit and achieve images which would otherwise be impossible they are clearly a powerful tool which when developed fully may allow previously impossible insights into the mircostructures at play within biological samples. The development and assessment of a aperture SNOM in combination with a IR-FEL for its use in biological studies will also be a considerable section of this thesis.

1.2 Barrett's oesophagus and oesophageal cancer

As cancer is such a large field it is obviously impossible to consider all its variants within a single thesis and therefore only cancer of the oesophagus will be studied. It is important to note that even though this thesis is aimed at oesophageal cancer, the instruments and analytical techniques demonstrated may be applicable to other forms of cancer or diseases. Oesophageal cancer was studied as it was the speciality of the clinicians who collaborated with the SCAnCan group, but it is a critical area of research as it has one of the fastest growing incident rates within the western world [63]. Although within the developed world oesophageal cancer is relatively rare compared to other cancers only affecting 0.45% of the population, it has one of the highest relative mortality rates at 0.35% [10], meaning that close to 80% of all patients diagnosed will die from it. The statistics are even worse for less developed areas of the world where the incident and mortality rates are 1.05% and 0.85% respectively. The high rate of mortality is generally due to the cancer undergoing spreading to other areas of the body before symptoms have become apparent leading to a diagnosis [64].

Often lifestyle influences such as alcohol consumption, smoking, dietary factors and obesity are attributed to a person developing oesophageal cancer but there can also be medical conditions which dramatically increase the likelihood of developing oesophageal cancer. An example of such a condition is Barrett's oesophagus (BO), which can increase the likelihood of developing oesophageal cancer by up to 40 times [65] and is therefore categorised as a precursor to cancer. Barrett's patients therefore have regular evaluations to assess if any early stage cancer is present within the oesophagus, so that treatment can be carried out rapidly, catching it before it spreads. This is usually done by taking biopsies of the oesophagus using an endoscope and implementing standard histology techniques to detect the presence of any early stage cancer, called dysplasia.

Barrett's oesophagus occurs when the lining of the oesophagus undergoes metaplasia, which is when the native cells within a region of tissue are replaced with cells of a different type. In the case of Barrett's oesophagus the normal lining of stratified squamous epithelium cells are replaced by columnar epithelium cells which have similar properties to that of the stomach lining [66]. This transformation is usually due to serious gastro-oesophageal reflux diseases, which results in the patient having chronic acid reflux. Acid reflux is when the stomach contents passes back into the oesophagus, the change in environment within the lower oesophagus is what drives the tissue to change as the cells are adapting to the new conditions. Hence why the metaplastic lining resembles that of the stomach rather than the oesophagus.

When tissue undergoes metaplasia it is prone to becoming dysplastic, which is to say that it has begun to appear abnormal or dysfunctional [67]. Tissue is described as being dysplastic when it has changed in both morphology and internal structures. Although not all occurrences of dysplasia will become cancerous it is known as being a pre-cancer indicator. Dysplastic tissues tend to be less structured and are often shown to have larger irregular nucleii. Previous work by members of the SCAnCan team used this abundance of nucleic material (DNA) to label cancerous tissue with a SNOM [68]. Dysplasia is such a strong indicator of the potential for cancer that Barrett's oesophagus patients who show clear dysplastic area in their biopsies will have surgery to remove the oesophageal areas along with and some of its surrounding.

There is a very limited amount of published work demonstrating the application of high spatial resolution techniques to oesophageal cancer. Previous work by Craig *et al* [69], showed that structural features found within Barrett's oesophagus samples called crypts can be imaged using a SNOM coupled with an IR-FEL. It was found that both the DNA and the glycoprotein signals were enhanced within the crypt region. At the same time the distribution of chemicals contributing to the Amide II signal were found to be anti correlated to the DNA. The spatial resolution shown in this study was capable of detailing the chemical microstructures within

the sample which would not be possible with techniques such as FTIR and Raman spectroscopy.

There are a greater number of far-field studies than near-field due to the larger number of such instruments being available [70–72]. In a study by Old *et al* [73], oesophageal cytology brushings were taken using an endoscope from patients with Barrett's and oesophageal cancer. A PCA-LDA classifier was used to label the collected cells which had been cytospun onto slides and imaged with FTIR spectroscopy. The generated classifier was shown to produce good sensitivity for Barrett's neoplasia at $\approx 80\%$ but poor specificity at only 60%. They were also not able to accurately classify nondysplastic Barretts oesophageal cells, which would be key for a diagnostic tool. This study is a proof of concept for IR imaging of diseased tissues, such as oesophageal cancer, to develop a new iteration of histopathology, but it also highlights that currently the methods are not ready for medical use.

In Timothy Craig's thesis [74], the application of ML techniques such as cluster analysis were applied to FTIR images of oesophageal tissue samples. This will be discussed in greater detail later in Chapter 3 as it led to the development of metric analysis, which is a newly developed ML algorithm. Within the thesis it is demonstrated that tissue discrimination is possible for oesophageal samples but it also shows that cluster analysis comes with limitations which may mean they are not ideal.

The aim of the research within this thesis is therefore to ascertain the potential strength of chemical imaging as both a diagnostic and research tool. A diagnostic tool should allow pathologists to assess if biopsies from Barrett's patients show any early signs of cancer while a research tool would be capable of supplying detailed information on the processes going on within the tissue.

Within this thesis samples of both tissue biopsies (Barrett's and oesophageal cancer) are studied along with commercial cell lines, which contain cells of only one specified type are studied. These samples are outlined in detail in the following chapter.

1.3 Thesis outline

Within this thesis there are two main areas of work, the first is the study of FTIR data using a newly developed algorithm to discriminate various sample types and to also provide potential insights into the chemical differences. The newly developed algorithm will be compared to random forest which is a commonly used and established machine learning technique. The second area of work is the development and application of an IR-SNOM instrument to image biological samples at resolutions not possible with other far-field spectroscopic instruments, to asses its usefulness for biological studies.

Chapter 1

Chapter 1 aims to give an introduction into cancer and the current state of cancer diagnostics. The current 'gold standard' for cancer diagnostics will be outlined along with the potential flaws involved. Chemical imaging is introduced with a brief introduction into the key fundamentals of how it may help with tissue diagnostics and its application for cancer. The need for machine learning is discussed with a focus on its application within this thesis. Barrett's oesophagus and oesophageal cancer is outlined as it is the focus of this thesis, with information presented as to the importance of studying this disease.

Chapter 2

Chapter 2 will mostly summarise the experimental techniques and instrumentation used within this thesis. A review of the general physics knowledge needed to understand the instrumentation is outlined along with the practises used to create and prepare the biological samples.

Chapter 3

In Chapter 3 the current state of machine learning for disease diagnostics will be discussed, with an overview of the work previous carried out within SCAnCan on ML. The methodology of the metrics analysis code which was developed to process and classify FTIR data is outlined using simplified artificially generated spectra for clarity.

Chapter 4

Chapter 4 will present the results of applying the metric analysis code to both cell lines and tissue samples. The performance of the metrics analysis code is compared to the random forest routine which is a commonly used machine learning process.

Within this chapter the first documented example of a machine learning algorithm distinguishing between cancer associated myofibroblasts and non-cancer associated myofibroblasts is demonstrated using metric analysis. This is currently very hard to do with standard optical techniques. The metric analysis was also shown to outperform an established ML technique in random forest, in both correct prediction rates and processing time. The metric analysis was tested on both tissue biopsies and cell line samples and was found to achieve a high performance in both studies.

Chapter 5

Chapter 5 outlines the work carried out with the SNOM and assess it performance and potential uses in cancer research. It will demonstrate the achievement of diffraction beating resolution and the potential implications to research.

In this chapter a spatial resolution of $0.15\,\mu m \pm 0.05\,\mu m$ was demonstrated which is 20 times better than the diffraction limit. A full image of an internal structure within the cell is presented, highlighting the potential of the SNOM. The

SNOM was compared to FTIR and was found to be considerably less topography dependent which is clearly beneficial.

Chapter 6

Chapter 6 will conclude the thesis highlighting the key achievements and predicting the possible future directions the research may continue in.

Chapter 2

Experimental Techniques

2.1 Microscopy and spectroscopy

Traditional microscopy uses the interaction of light, usually light from the visible region of the electromagnetic spectrum (EM), with the sample to generate contrast in the generated image. At visible wavelengths the contrast is mostly due to the morphology of the sample and contains very little information about it's chemical composition. To overcome this limitation many stains have been developed which act to highlight the presence of important biological molecules such as proteins, lipids and nucleic acids. The most commonly used stain is Hematoxylin and Eosin (H&E) which colours areas which are heavy in proteins pink and areas rich in nucleic acids blue. H&E staining and modern microscopes have become the standard around the world for differentiating tissues by capitalising on the added contrast given by the stain which gives considerably more insight into the chemical distribution within the sample. The main limitations of modern microscopy for tissue discrimination is the time required to prepare the samples and then manually study them under a microscope. There is also the added issue that this methodology is somewhat subjective as it is up to the histologist to determine the result based on the apparent tissue features.

Due to these limitations in traditional microscopy there has been a drive for quicker and more quantitative methodologies for tissue discrimination, which with the recent advancements in computational technology and instrumentation has seen infrared (IR) spectroscopy emerge as a front runner. Spectroscopy relies on the unique variation in absorbance over a range of IR wavelengths for a given sample. IR light is absorbed due to the excitation of the vibrational states of the molecules within the sample. As different areas of a tissue samples will contain varying amounts of the critical biological molecules there will a notable variation in the collected spectra. By studying these spectral differences it is possible to distinguish a large array of different tissue types. IR spectroscopy has become prevalent compared to other wavelength regions due to the photon energy within the IR range being equal to that of the vibrational transitions of many common biological molecules. IR spectroscopy is therefore a technique which is sensitive to the chemical composition within the sample allowing for the distribution of such chemicals to be imaged without any dyes. Spectra are traditionally displayed in terms of wavelength, which is the inverse of the wavenumber. The key strength to spectroscopy is that it is able to give insight not only on the morphology but also the chemical content within the sample, which gives the classifier more information to distinguish the tissue types.

2.2 Physics of optics

The physics of how light interacts with materials is critical to the understanding of the instruments used within this thesis and as many principles will be used repeatedly throughout, this section aims to detail the relevant physics ideas.

2.2.1 Diffraction limit

A central principle within optics is the diffraction limit, which defines the smallest resolvable object a given instrument is able to image. The term diffraction limit

comes from the fundamental restraint on the resolution imposed due to far-field diffraction effects. This limit has been known for some time, as Ernst Abbe in 1873 showed that the best resolution is approximately half that of the wavelength using visible light [75]. The full equation which defines the smallest resolvable object is:

$$d_{limit} = \frac{\lambda}{2nsin(\theta)} \tag{2.1}$$

where d_{limit} is the size of the smallest resolvable object, λ is the wavelength of the light, n is the real component of the refractive index of the material which the light is travelling through and θ is the objective lens half aperture angle. Since in air $n \approx 1$ and $\theta = \frac{\pi}{2}$ for an ideal lens, Equation 2.1 simplifies to:

$$d_{limit} = \frac{\lambda}{2} \tag{2.2}$$

For the mid IR wavelengths of 5-10 µm used for the experiments within this thesis the resolution would be limited to 2.5-5 µm. The observed resolution may be larger though as these are values under ideal conditions and things such as misalignment and aberrations due to the optics which harm the resolution.

2.2.2 Wave behaviour at a boundary

When light reaches a boundary between two materials, such as when the light is exiting a sample and passing into the air, it is capable of both reflection and transmission. An important material property which determines how the light will behave is a medium's complex refractive index $\underline{n}$. A complex refractive index is made up of two parts as shown in Equation 2.3, where n is the real component and the l is the imaginary component.

$$\underline{n} = n + il \tag{2.3}$$

The real component of a material's refractive index is given by the following equation where c_v is the speed of light in a true vacuum and c_m is the speed of light in

the material.

$$n = \frac{c_v}{c_m} \tag{2.4}$$

Depending on the angle of incidence light will undergo varying amounts of reflection and transmission at the boundary. Figure 2.1 shows that the transmitted wave will have changed direction and therefore has undergone refraction. The angle of the transmitted beam from the normal can be calculated using Snell's law shown below:

$$n_1 sin(\theta_i) = n_2 sin(\theta_t) \tag{2.5}$$

where n_1 and n_2 are the real components of the respective materials refractive index and θ_i is the angle of incidence and θ_t is the angle of refraction.

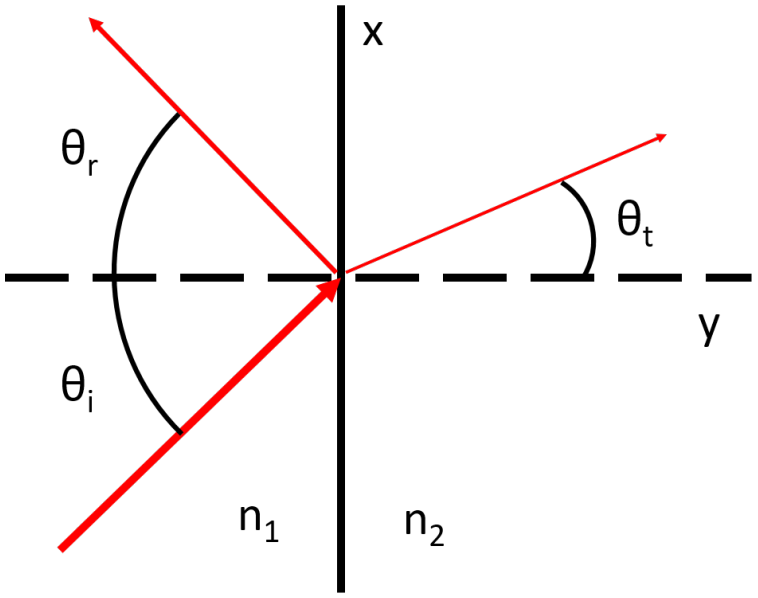

FIGURE 2.1: Diagram showing the interaction of light at a boundary between two materials with different refractive indices.

Snell's law can be used to demonstrate a situation where all the light is reflected back into the medium. This phenomenon is called total internal reflection and occurs when $n_1 > n_2$ and the angle of refraction is $\geqslant 90°$. The critical angle θ_c defines the minimum incident angle which will facilitate total internal reflection and can be determined by rearranging Snell's law after having $\theta_t = 90°$, the

equation for the critical angle is shown below:

$$\theta_c = sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{2.6}$$

Total internal reflection is the main principle of how optical fibres, such as the one used in the SNOM instrument which is discussed later, are able to efficiently transmit light along a non linear path.

Fresnel's equations can be used to define the proportion of light which will be reflected at a boundary by calculating the reflection coefficient for both s and p polarised, which are shown below:

$$R_s = \left(\frac{n_1 cos(\theta_i) - n_2 cos(\theta_t)}{n_1 cos(\theta_i) + n_2 cos(\theta_t)}\right)^2 \tag{2.7}$$

$$R_p = \left(\frac{n_1 cos(\theta_t) - n_2 cos(\theta_i)}{n_1 cos(\theta_t) + n_2 cos(\theta_i)}\right)^2 \tag{2.8}$$

where R_s and R_p are the reflection coefficients for s and p polarised light respectively. Since the transmission coefficients and reflection coefficients sum to 1, the proportion of light which is transmitted across a boundary can be simply found by:

$$T = 1 - R \tag{2.9}$$

Another important phenomenon which is a fundamental principle behind many instruments such as attenuated total reflectance FTIR and SNOM is the creation of evanescent waves. The existence of evanescent waves can be shown by building on the equations already shown. The derivation below is for the generation of evanescent waves within an attenuated total reflectance FTIR instrument and not that of a SNOM as the near-fields produced within a complex sample such as tissues is currently too complicated to clearly define mathematically. The result below does ultimately demonstrate the general structure of near-field radiation which is that of a exponential field that decays very quickly, which is common for all near-fields independent of their generation.

As previously stated, if the angle of incidence is larger than the critical angle total internal reflection will take place. But as will be shown this isn't the whole picture.

If Snell's law is rearranged and applied to the condition for total internal reflection which is $\theta_i > \theta_c$ the result is Equation 2.10

$$sin^{-1}\Big(\frac{n_2}{n_1}sin(\theta_t)\Big) > \theta_c \tag{2.10}$$

By substituting in the equation for the critical angle and simplifying we get a constraint on θ_t when undergoing total internal reflection.

$$sin(\theta_t) > 1 \tag{2.11}$$

This constraint can be used to show how evanescent waves are formed by applying it to the equation for the transmitted plane wave electric field equation $\vec{E}_T$ shown below:

$$\vec{E}_T = \vec{E}_{0T}e^{i(\vec{k}\vec{x}-\omega t)} \tag{2.12}$$

where $\vec{E}_{0T}$ is the amplitude, $\vec{k}$ is the wave vector, $\vec{x}$ is the position vector, ω is the frequency of the light and t is time. In Equation 2.12 the wave vector and position vector can be expanded, giving them in terms of the x and y axes. For this example the system will be a 2D plane as shown in Figure 2.1.

$$\vec{E}_T = \vec{E}_{0T}e^{i(k_x x+k_y y-\omega t)} \tag{2.13}$$

The equation can then be altered so that it is in terms of k_T rather than k_x and k_y, which gives:

$$\vec{E}_T = \vec{E}_{0T}e^{ix(k_T sin(\theta_T))}e^{iy(k_T cos(\theta_T))}e^{-i\omega t} \tag{2.14}$$

The important term in Equation 2.14 which helps to show the properties of the evanescent waves is the second exponential which describes the field in the y-axis. First the cosine function has to be converted into a sine function which can be

done with the following equation.

$$cos(\theta_T) = \sqrt{1 - sin(\theta_T)^2} \tag{2.15}$$

It was shown in Equation 2.11 that $sin(\theta_T) > 1$ which by using Equation 2.15 shows that $cos(\theta_T)$ must a complex term. Equation 2.14 therefore becomes:

$$\vec{E_T} = \vec{E}_{0T} e^{ix(k_T sin(\theta_T))} e^{-Ay} e^{-i\omega t} \tag{2.16}$$

where A is an arbitrary number. The real component of Equation 2.16 can be found which depicts the structure of the evanescent wave present in the second medium, which is shown below.

$$Re(\vec{E_T}) = \vec{E}_{0T} cos(x k_T sin(\theta_T) - \omega t) e^{-Ay} \tag{2.17}$$

Equation 2.17 shows that on the surface of the original medium, which in the case of SNOM would be the sample, is a oscillating non-propagating wave which decays exponentially the further away from the samples surface. Evanescent waves are often very hard to detect as they only extend a very small distance from the surface. As previously mentioned evanescent waves are critical to the operation of scanning near-field optical microscopy which will be discussed in detail later.

2.2.3 Absorption

Absorption is integral to the studies carried out within this thesis as it is the foundation of spectroscopy and chemical mapping. Previously it was shown that the behaviour of light at a boundary between two mediums is dependent on the real component of the refractive index of each material. Absorption instead details how the light interacts within a given medium. The absorbance of a given material A, can be calculated by using:

$$A = \frac{4\pi l L}{\lambda} \tag{2.18}$$

where l is the imaginary component of the materials refractive index, L is the path length and λ is the wavelength of the light. The importance of Equation 2.18 is that it shows absorbance is dependent on both the refractive index of the material determined by chemicals within and also the wavelength of the light used to probe it.

Another commonly used relation within spectroscopy is the Beers-Lambert shown in equation 2.19:

$$A = \epsilon L c \tag{2.19}$$

where A is absorbance, ϵ is the molar absorptivity coefficient, L is again the path length and c is the analyte concentration. The molar absorptivity coefficient is dependent on both the molecule in question and the wavelength of the probing light. This is a useful relation as it demonstrates the intuitive relationship between the sample composition and the absorbance. As both L and c essentially describe how much of the molecule is present within the sample and ϵ describes the likelihood of absorption occurring depending on the wavelength of the incident light.

2.3 Fourier transform infrared spectroscopy

Fourier transform spectroscopy (FTS) was first developed in the late 60's and has been used to gather spectral information in a wide variety of applications [11, 76–78]. But because of the limited computational throughput of the computers of the time the development was slow. Two decades after its first introduction advances in computer power enabled FTS techniques since they were better equipped to handle, what was at the time, large computational requirements needed to carry out the Fourier transform calculations. This caused the field of FTS to bloom with many variations theorised and tested, including visible and ultra-violet wavelengths sources [79–81] but the most successful and now commonly used is Fourier transform infrared (FTIR) spectrometry. The strength of FTIR is that it has the potential to excel in biological research due to the many organic chemical absorption features present within the infrared region of the electromagnetic spectrum. It is for this reason that FTIR is one of the techniques at the forefront of research for tissue diagnostics.

All the FTIR data presented in this thesis was collected on a spectrometer which is part of Professor Peter Gardener's group at the Manchester Institute of Biotechnology. The machine is a commercial instrument developed by Varian, shown in Figure 2.2, and consists of a Varian Cary 620 FTIR microscope coupled with a Varian Cary 670 spectrometer, with a broadband GloBar IR light source. Varian has since been bought by Agilent Technologies, Santa Clara CA, USA. The instrument is fitted with a 128x128 pixel N_2(liq) cooled mercury-cadmium-telluride (MCT) focal plane array (FPA) detector. By using an FPA rather than a single sensor detector as used in other FTIR instruments, it is possible to collect 16,384 spectra from various spatial positions on the sample simultaneously. This is one of the major strengths of FTIR in that many absorption spectra can be collected over a wide IR range in a relatively short amount of time. For the vast majority of images taken the FITR was setup to have a pixel size of 5.5 µm giving an image size of 704 µm x 704 µm. A higher magnification condenser can be used to decrease the pixel size to 1.1 µm and although this technically allow for smaller pixel sizes

resulting in higher detail, it doesn't mean the resolution has become 1.1 µm as FTIR is still limited by far-field physics. This is the major limitation of FTIR in that it does not manage to overcome the diffraction limit discussed previously and therefore can struggle to image the very fine details within a sample.

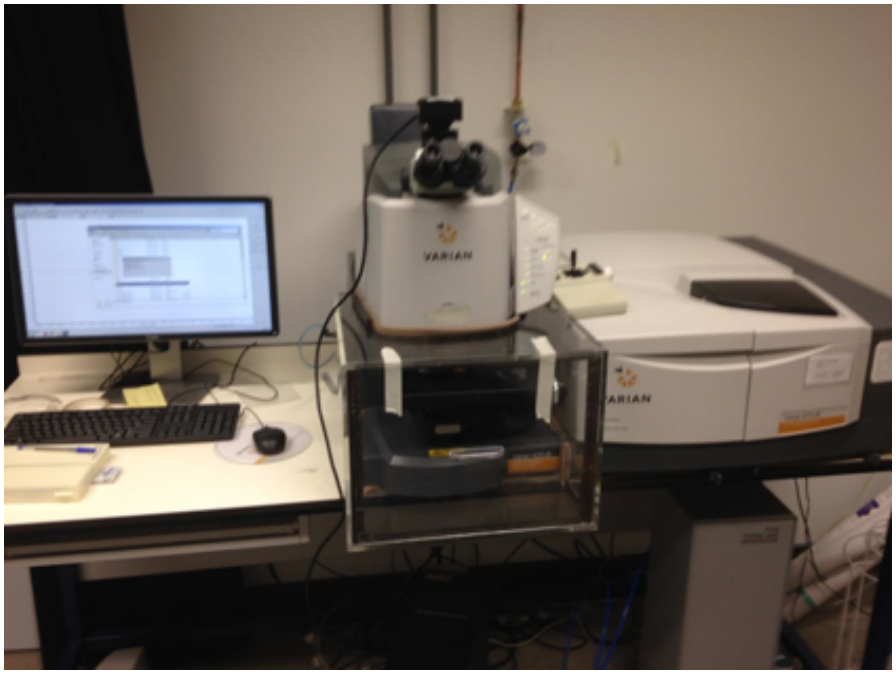

FIGURE 2.2: Professor Peter Gardner's FTIR instrument at the Manchester Institute of Biotechnology.

The sample is placed within a purge box which creates and maintains an environment of dry air with low levels of CO_2 around the sample. A purge gas generator feeds a constant flow of purged air into a loosely air tight perspex box which surrounds the sample stage, resulting in a higher than atmospheric pressure within the FTIR, which acts to stop any of the ambient air from entering the instrument. This is important since both CO_2 and H_2O have large absorption features within the same IR spectral range that is significant for biological studies. A humidity monitor is used within the purge box to measure the H_2O content of the air and every time the perspex box is opened to replace a sample approximately 15-20 minutes is needed before scanning for the levels to drop to an acceptable level.

A background scan is needed to remove features in the spectra due to, contaminants on the optics, variations in the H_2O content of the air, absorption within the slide or differences in the pixel efficiencies on the FPA. The background is taken on a clean slide to prevent image artefacts due to contaminants being present in the background image. This has to be rerun every time the purge box has been opened and the H_2O has stabilised.

2.3.1 FTIR principles

In most FTIR spectrometry equipment a broadband thermal IR light source is used to illuminate the sample, but with recent developments in tunable IR lasers such as quantum cascade lasers, new light sources have begun to be used within various microscopy and spectroscopic instruments [82, 83]. There are multiple modes of operation used in modern FTIR, transmission, reflectance and attenuated total reflection (ATR). The principles of how each work are similar only varying in how the light illuminates the sample, Figure 2.3 demonstrates these differences. In transmission mode, which is exclusively used for the work in this thesis, the IR light passes through both the sample and the IR transmissive slide it is mounted on. In reflectance mode the light strikes the sample at an angle and reflects off the IR reflective slide on which the sample is mounted. Finally, in ATR mode a IR transmissive prism is used to mount the sample. The light internally reflects off the boundary between the sample and prism creating evanescent waves which penetrate into the sample. The evanescent waves are only able to probe a small distance in to the sample, usually only around a micron, allowing for a different sample thickness to be studied compared to transmission and reflectance.

After the light has interacted with the sample via any of these modes it is then passed to a interferometer before finally reaching the detector. For most modern infrared spectrometers a Michelson interferometer is used, a schematic of a Michelson interferometer is shown in Figure 2.4. A Michelson interferometer works by equally splitting the incoming light from the sample into two arms using

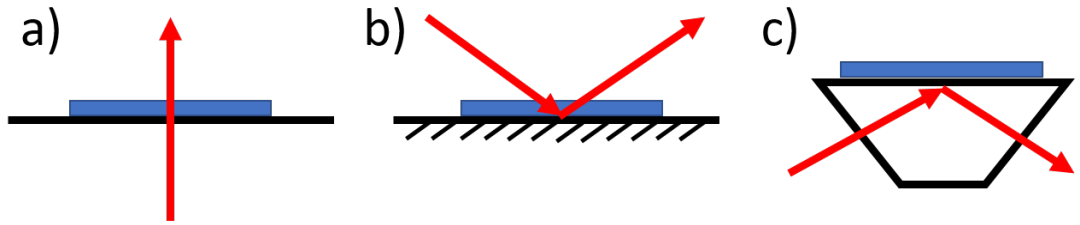

FIGURE 2.3: A diagram showing the various FTIR imaging modes, a) Transmission mode where the light passes through the sample and the transmissive slide, b) Reflectance mode which shows the light passing through the sample where it then reflects off the IR reflective slide and c) Attenuated total reflection (ATR) mode which uses a prism made of IR transparent material to create evanescent waves at the sample-prism boundary to probe the sample.

a beam splitter, which is usually made of CaF_2 or near infrared (NIR) quartz for near IR instruments. One arm is sent to a fixed IR mirror while the other travels to a computer controlled oscillating IR mirror, which moves in the axis of the beams direction. The light reflects off both mirrors recombining back at the beam splitter where they will undergo constructive and destructive interference. The interference effect will vary depending on the different frequencies of light which make up each beam, therefore encoding the spectral information within the time domain of the recombined beam.

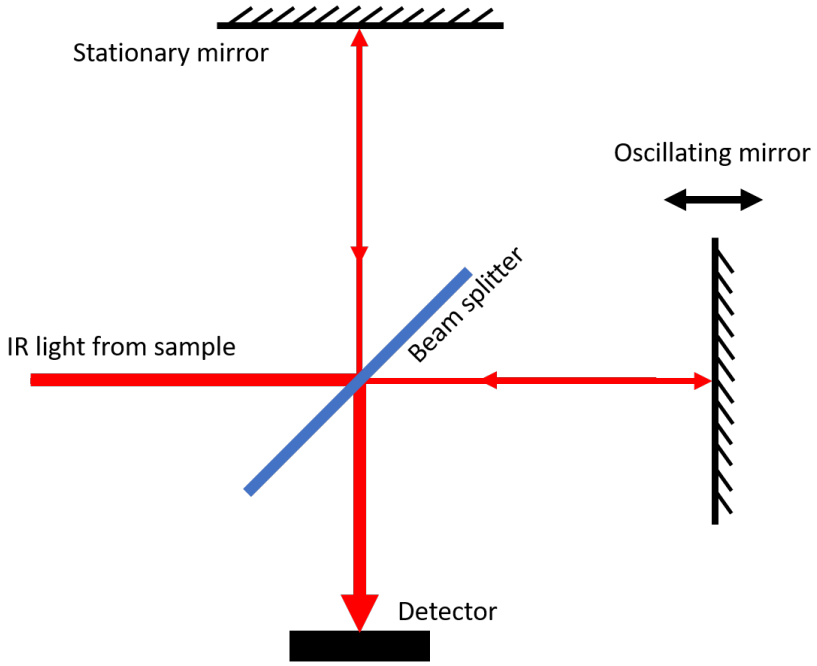

FIGURE 2.4: Schematic of a basic Michelson interferometer.

The recombined beam is then recorded by an IR FPA detector as a interference pattern, called a interferogram. An example of an interferogram taken from an image of Barrett's oesophagus tissue is shown in Figure 2.5. To convert the spectral information from the time domain $f(t)$ into the frequency domain $f(v)$ a Fast Fourier transform (FFT) is used, shown in Equation 2.20.

$$f(v) = \int_{-\infty}^{\infty} f(t) e^{2vi\pi t} dt \tag{2.20}$$

Once the FFT has been applied to the interferogram, the result is a spectrum showing the absorbance of the sample over a range of wavenumbers, Figure 2.5 shows the result of performing an FFT on the interferogram. The benefit of using an interferometry spectrometer is that the data gathered at every increment measured by the detector contains information over a large spectral range. Whereas standard grating spectrometry only gathers information over a very narrow band of the spectrum given by the size of the grating.

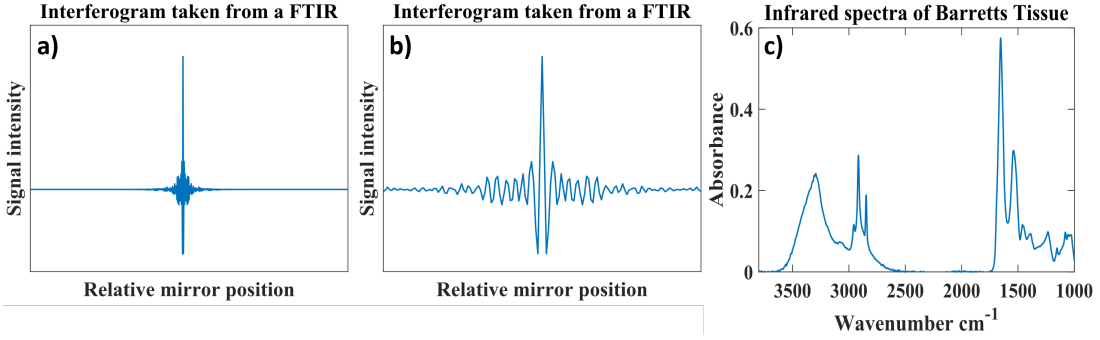

FIGURE 2.5: A typical interferogram and absorption spectra taken from a FTIR image of Barrett's oesophagus tissue. In a) a full interferogram, b) a cropped section from a) to better show the interference pattern and c) an absorption spectra formed by carrying out a FFT on the interferogram shown in a).

FTIR also benefits from using interferometry due to the multiplex advantage, originally proposed by Fellgett, which acts to improve the signal to noise of the data collected [84]. In instruments such as FTIR a very broad light source is used and hence a large amount of the light is incident on the sample/detector. An alternative method to interferometry is to use a variable grating to sweep the sources wavelength range in a number of discrete steps, with each allowing only a

very narrow wavelength range to reach the sample/detector. This would result in a considerably lower intensity at each step as the grating simply blocks the unwanted wavelengths, which produces a much weaker signal produced by the detector. Interferometry is therefore able to record a much larger signal resulting in an increased signal to noise on the detector output. This advantage is compounded as a large component of the noise in IR experiments is often attributed to the detector, unlike at other wavelength where factors such as the source and optics dominate the noise, which means the quality of interferometry data is often superior. The multiplex advantage is therefore one of the reasons why IR interferometry such as FTIR has flourished. To further improve the quality of the data taken each FTIR image is made up of 256 individual scans which are averaged together.

2.3.2 FTIR data structure

The detector on the FTIR used at the Manchester Institute of Biotechnology uses a IR FPA which means it has a 2D array of IR sensors rather than just a single one as used in some FTIR instruments. It is therefore possible to take multiple spectra at different spatial positions on the sample simultaneously. The data generated by the FTIR is therefore a 3D hypercube with the dimensions of 128x128xn where n is the number of wavenumbers recorded in the spectra. An image representing a hypercube is shown in Figure 2.6, where the x and y-axes show the variation in absorption spatially and the z axis represents the variation at different wavenumbers. A hypercube can be envisioned as many images stacked on top of each other with each image showing the variation in absorption spatially at a different wavenumber.

Although the number of spectra within an image is fixed by the number of sensors in the FPA, a computer controlled motorised stage can be used to take multiple images at different areas of the sample. These images can then be stitched together post scan to create a larger mosaic image.

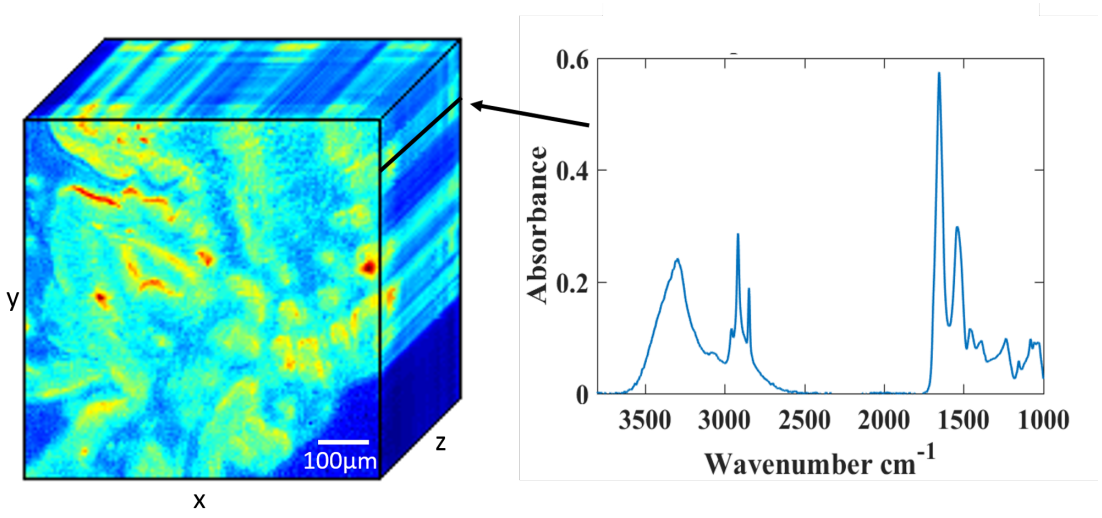

FIGURE 2.6: A 3D representation of a hypercube taken of Barrett's oesophagus tissue. The 2D face shows the spatial variation in absorption at a single wavenumber. Whereas the z axis shows the variation in absorption at one point on the sample over many wavenumbers, producing a spectra such as the one shown.

2.4 Infrared scanning near-field optical microscopy

Standard optical microscopy at visible wavelengths has been widely used in biology and medicine all around the world for hundreds of years. It has been used in both research and as a tool for studying a vast range of samples. This is partly due to its relatively low cost, ease of use and powerful capabilities for gaining critical diagnostic information. Given these advantages it is still limited by the same far-field physics shown previously to restrict FTIR. For visible wavelengths of $\lambda \approx$ 0.4-0.8 µm the theoretical optimal resolution is around 0.2-0.4 µm and in reality is often worse due to the imperfect optics used. It is this limit in resolution and the desire to image smaller and smaller objects that has driven the development of techniques which aim to improve the resolution of imaging instruments.

One method to improve resolution would be to use smaller wavelengths as the probing light, since the resolution is proportional to the lights wavelength. This would improve the resolution but isn't always appropriate since it can often be difficult to produce the necessary optics to efficiently work at the given wavelength. There is also the added complication that different areas of the EM spectrum will

result in different absorption features being present due to the changing mechanisms of interaction between the light and the sample, which may make the images hard to interpret. Also just because a light source of smaller wavelength can be used to gain better spatial resolution it doesn't necessarily mean that the absorption spectra will contain any of the desired information.

An extension to the thought of using smaller wavelengths are techniques such as scanning electron microscopy (SEM) and transmission electron microscopy (TEM). Electron microscopy uses energetic electron beams to act as a source of very short wavelength light to give incredibly high spatial nanometre resolution. The wavelength (λ) of the electron beam can be calculated by using the relativistic corrected de Broglie condition shown in Equation 2.21:

$$\lambda \approx \frac{h}{\sqrt{2Em_0(1+\frac{E}{2m_0c^2})}} \tag{2.21}$$

where h is Planck's constant, E is the accelerating energy on the electron beam, m_0 is the rest mass of an electron and c is the speed of light in a vacuum. There are limitations with electron microscopy techniques for this area of research, chiefly being that the samples must be placed within a high vacuum which is often not possible for biological samples. There are potential solutions to this as many steps can be taken to prepare the sample such as chemical fixing, dehydration and embedding may make it capable of surviving the vacuum but this is often not guaranteed and these are lengthy and laborious tasks which increase the preparation requirements [85]. The high intensity of an electron beam require to image the tissue samples would tend to damage it as it interacted with the sample. A further complication of imaging biological samples such as tissue is that even when dried they are relatively thick at a few microns. This is much larger than the usual sample thickness of $< 100\,\text{nm}$ which the electron microscopes are designed for. Electron microscopes such as TEM and SEM may also not provide the desired chemical information needed for the purposes of tissue differentiation as they are often focused on the morphology of the sample rather than the chemical content. There are methods such as electron energy loss spectroscopy (EELS) which may be

more appropriate but it still suffers from many of the limitations present in TEM and SEM. EELS is capable of assessing the presence of key elements within the sample by measuring the inelastic scattering of the electrons. This is not so helpful for tissue differentiation as the key is the variations in the complex molecules such as proteins which distinguish the two and not the elements which are common to most biological substances. The need therefore for an in air technique able to beat the diffraction limit with IR light becomes apparent.

2.4.1 SNOM principles

The first outline for an instrument able to overcome the diffraction limit was presented as far back as 1928 by Synge [86]. He stated that a very small aperture illuminated from behind could be used as a subwavelength sized light source. The aperture would then be placed very close to the samples surface so that the distance between the two was much less than the wavelength of the light. This would result in the collected light having only interacted with a very small volume of the sample directly beneath the aperture. The resolution of such a device is therefore proportional to the size of the aperture rather than the wavelength of the light, which is it's key strength since high resolution imaging is possible without the need for a diffraction limited lenses to focus the light. An added benefit of the resolution being independent of the probing light's wavelength is that it allows for the high-resolution imaging at any wavelength with desirable absorption characteristics, such as IR. Near-field optical microscopy would therefore be capable imaging biological samples using infrared wavelength light at subwavelength resolution. Even in the early stages of near-field microscopy's development the technique was envisioned to help with biological problems, by capitalising on the high spatial resolution capabilities which other methods could not [87].

Although the principles of aperture near-field imaging were first proposed almost a century ago, due to the complex technology required for such a device it was not demonstrated experimentally until 1972 by Ash and Nicholls [88]. They used a microwave source of $\lambda \approx 3$ cm, to gain a resolution of around $\lambda/60$, which

was the first experimental evidence that near-field optical instruments were capable of overcoming the diffraction limit, beating it by a factor of 30. Since the separation of the sample and aperture must be very small, the mechanical requirements needed to advance near-field imaging to visible or infrared light sources becomes much more substantial. These requirements hampered the development of near-field imaging at submicrowave wavelengths until the development of scanning tunnelling microscopy (STM) in 1982 by Binnig and Rohrer [89].

STM was the first imaging technique to successfully use the revolutionary new piezoelectric drives capable of controlled movements over nanometre distances. The introduction of these drives led to an explosion in the field of scanning probe microscopy (SPM), due the high precision capabilities of piezoelectric drives to set and maintain very small probe to sample distances. Piezoelectric drives can also be used to move the sample very small distances underneath the tip allowing for the precise scanning of very small image areas not previously possible. It was not long after the development of STM that the same technology was applied to near-field optics by Phol *et al* in 1984 [56]. They were able the obtain the first diffraction breaking resolution images with visible light achieving a resolution of $\lambda/20$ ($\approx$ 25 nm) by using an optical fibre as a very small aperture to carryout the near-field imaging. This technique was called scanning near-field optical microscopy (SNOM/NSOM). The premise of this technique is to capitalise on the evanescent waves present at the samples surface. In the immediate vicinity of an object radiating electromagnetic waves there is a near-field component that consists of high frequency contributions that contain information on the physical structure of the object at length scales shorter than the wavelength of the EM wave. This near-field contribution decays rapidly with distance from the source. However an aperture within the near-field region will capture these high frequency contributions and yield information on length scales shorter than the diffraction limit. Since the aperture is very small only light from a small area of sample is able to propagate into the fibre. Finally near-field optics began to gain traction as a powerful technique capable of gaining important chemical information common in many

standard optical techniques but at vastly improved resolutions. Many advancements have been made within near-field optical microscopy, with other techniques finding success in overcoming the diffraction limit such as, tip enhanced Raman spectroscopy (TERS) [90, 91] and near-field optical random mapping (NORM) microscopy [92]. This thesis though will focus on SNOM as used by the SCAnCan group for subwavelength resolution near-field imaging.

2.4.2 SNOM instrumentation

Modern SNOMs generally work, as pioneered by Pohl *et al*, by using an optical fibre to create a subwavelength sized aperture capable of being scanned over the samples surface. To obtain a very small aperture a fibre with an inner core diameter of the desired aperture size can be used, but this isn't always possible as manufacturing a reliable fibre with an inner core with a submicron diameter would be extremely hard. It is possible though to create a submicron aperture by sharpening one end of the fibre into a tip which is then coated with an opaque metal in such a way to create a tiny aperture over the fibres core, this will be discussed in detail later in the chapter. The principles of SNOM are applicable to any wavelength of light but in practice an appropriate fibre which allows efficient transmission of the light isn't always currently available. For most IR instruments and the SNOM used within this thesis fibres with a core of selenide glass was used. An example of a IR SNOM fibre is shown below in Figure 2.7.

The SNOM tip is brought very close to the samples surface using piezoelectric drives and a feedback mechanism is implemented to maintain the separation between the two, discussed in detail later in this chapter. This is the general methodology common to most SNOM instruments but there are some variations in configuration between instruments, which are outlined in Figure 2.8.

The most common variations are in how the light is incident on the sample and how the fibre is utilised. As Figure 2.8 shows the fibre can be used as a subwavelength light source and is used to illuminated the sample. Instead it can

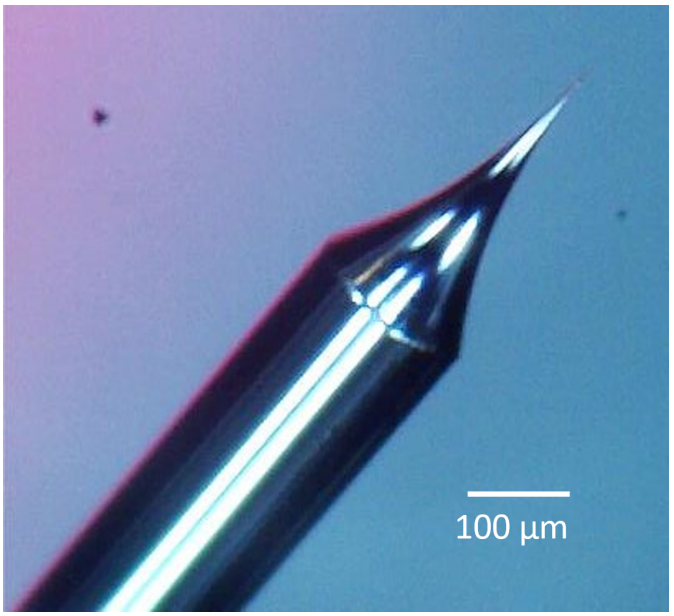

FIGURE 2.7: An example of an etched infrared fibre used on the SNOM.

be used as a collector to capture the light over a very small area on the samples surface. In the illumination configuration the light source is directly coupled to the non-tip end of fibre, where the light then propagates along the fibre until it exits via the aperture. The light passes through the small volume of sample underneath the tip where it will ultimately be collected by a detector. Alternately a SNOM can be set up so that the light directly from the light source strikes the sample underneath the fibres tip where it then enters via the aperture and thus the fibre acts as a collector instead of a light source. In this variation, the light propagates along the fibre before exiting the end of the fibre which is positioned directly in front of the detector. In both configurations the light recorded by the detector has only interacted with the a small amount of sample under the aperture and therefore only contains the chemical information of a small volume of sample. The second commonly found variation is the angle at which the light is directed/collected beneath the fibres tip. In transmission mode the light is collected/directed at 90° to the sample, which means the light travels through the entire sample and the IR transmissive slide. In reflection mode the light is instead collected/directed at a shallow angle of $\approx 15°$ to the sample. The SCAnCan group's current iteration of the SNOM uses the fibre as a collector and can operate in both modes since the optics are able to redirect the light along either path. Due to the different possible

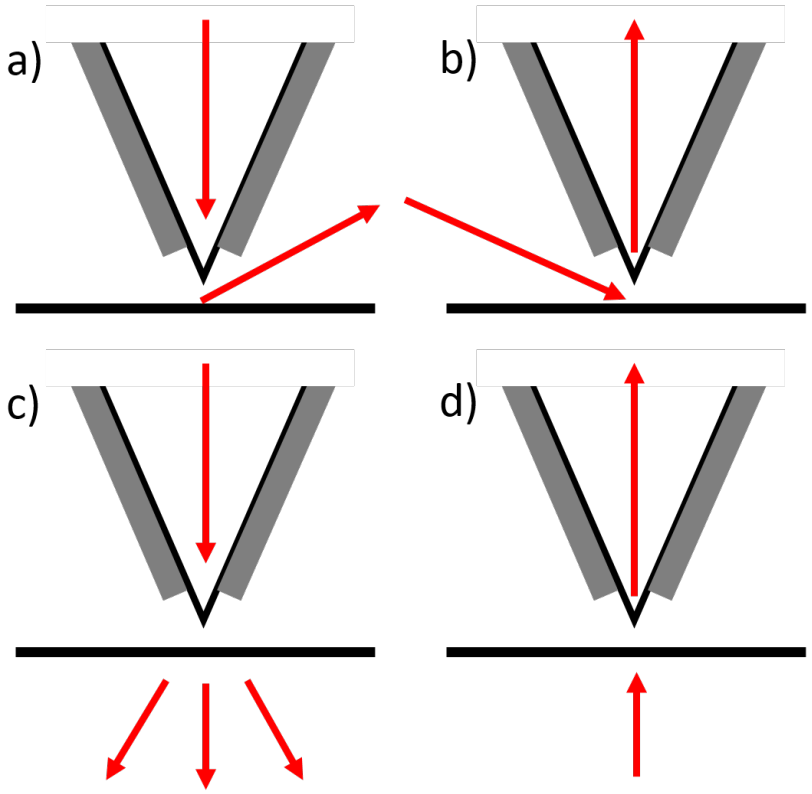

FIGURE 2.8: A diagram showing the different operational configurations commonly used in SNOM. a) Illumination via tip in reflection mode; b) Collection by the tip in reflection mode; c) Illumination via the tip in transmission mode; d) Collection by the tip in transmission mode.

penetration depths in both versions the two modes may have insight into different information and therefore produce different images which will be explored in a later chapter.

Another SNOM variation not used by the SCAnCan group but which is worth mentioning is fluorescence SNOM [93–95]. Instead of using a light source to illuminate the sample it uses fluorescent molecules within the sample to act as a light source. The SNOM tip is rastered over the sample where it is able to detect the emitted light from the fluorescent molecules and therefore map their spatial distribution. Very high resolution images have been produced using such techniques even being able to image molecules [96]. This like most techniques has drawbacks such needing the sample to be prepared with fluorescent labels which aren't feasible for all sample types and is an added stage needed in sample preparation. It may also be only possible to prepare a sample with a single label and therefore may

only give insight into the distribution of one chemical signature, unlike techniques such as illuminated SNOM and FTIR which can image the same sample at many wavelengths producing data that may contain much more chemical information giving a greater insight.

Throughout the SCAnCan group's work with SNOMs there have been many iterations of the SNOM instruments which over the last 7 years have been continuously upgraded with additions and whole new machines developed, with the aim to improve the capabilities of the instrument and quality of the images collected. The original instrument was a much smaller and simpler version of the current iteration and was only able to image in reflection mode. The current SNOM is a larger machine capable of imaging in both modes and has additional features such as an inverted optical microscope and a larger sample stage not present on older instruments. For the sake of brevity and focus this thesis will not outline the full development of the SNOM instruments used by the SCAnCan group in detail since this has already been done by Timothy Craig in his thesis [74] and will instead it will focus only on the current iteration used to produce the data shown within the thesis.

The SNOM images presented within this thesis are from a non-commercial instrument originally built by Antonio Cricenti and Marco Luce [97]. The instrument has since been repeatedly improved by both the original designers and other members of the SCAnCan group. The SNOM was located at the Scientific and Technology Facilities Council (STFC) Daresbury Laboratory which is in the North West of England. The SNOM was coupled with a high power tuneable infrared free electron laser (IR-FEL) attached to the ALICE accelerator (Accelerators and Lasers In Combined Experiments) [98]. SNOM experiments have been undertaken at Daresbury throughout the years of 2011-2017 in one form or another. Due to the operational requirements of ALICE the SNOM was used for periods of around 2-4 months each year, where it mostly ran 24 hours a day for 5-7 days a week. It was between these scheduled periods of accelerator time that the SNOM and the IR light source were upgraded for the next run. The following sections will

break down the full configuration of the current SNOM into sections describing each element in further detail.

2.4.3 SNOM in conjunction with IR-FEL

For the majority of the time the SNOM was in use it was in conjunction with an IR-FEL. FELs were first developed in the 1970's by a group at Stanford University [99] and have been growing in popularity ever since. A FEL is an excellent choice for a IR light source to use with the SNOM since it is able to produce a very intense source of IR light over a broad tuneable range of 5.5-9 µm. This is ideal for the SNOM since the size of the aperture results in a very small percentage of the light being captured and so having an intense source allows for a larger signal to be recorded by the detector. The major limitation of FELs are the complexity and cost of building and maintaining such a device since they require an accelerated electron source and all the equipment which that entails.

The SNOM is located in the diagnostic room which is situated next to the ALICE accelerator hall to minimise the distance the IR light needs to travel in the beamline from FEL to SNOM. The diagnostic room has adequate shielding so that personnel can operate the SNOM while the ALICE accelerator is running, which is why the SNOM is not within the accelerator hall. The SNOM and its optics are placed on a heavy optics table, as shown in Figure 2.9, to minimise any vibrations which the SNOM is very sensitive too.

As seen in Figure 2.9 the SNOM has an enclosure which can be shut while scanning to both block external IR radiation from interfering with the SNOM imaging but it also protects the personnel working in the room from any potentially harmful IR radiation. It can also be seen that the majority of the electronics used to operate the SNOM are not on the table and are instead housed in an electronics rack which frees up space on the table and also prevents the electronics having a detrimental effect on the SNOM such as vibrations from cooling fans.

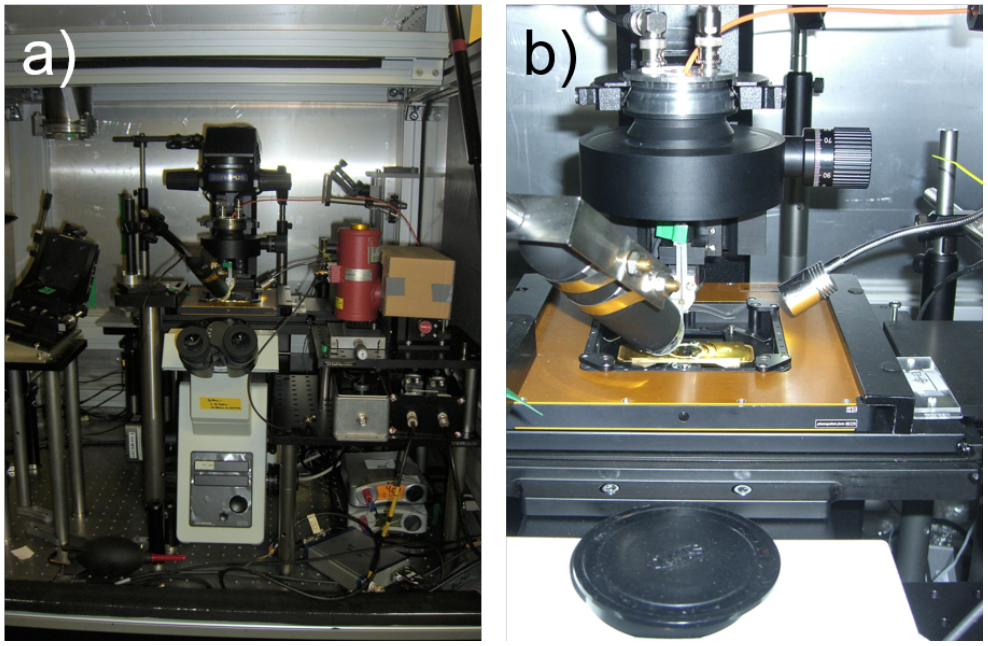

FIGURE 2.9: Pictures taken of the SCAnCan group's SNOM, a) shows the whole SNOM and b) shows the piezostage sample holder and SNOM head in greater detail.

2.4.3.1 FEL light source

ALICE and the attached FEL is an example of a fourth generation light source, a detailed diagram of both is shown in Figure 2.10. Light sources have historically been grouped into generations with each new generation outclassing the previous by orders of magnitude in areas such as brightness, coherence or pulse duration [100], with the fourth generation currently being the newest. ALICE is also a energy recovery linac, which means it is able to recover a large amount of the energy from the previously accelerated electron bunch which can be in turn be used to accelerate the following bunch [101]. This is possible because once an electron bunch has been used by the FEL it re-enters the linac out of phase with the RF field inside and therefore allows for it's kinetic energy to be recovered. By recovering the energy from each bunch the linac is able to produce high energy electron bunches very efficiently, which in turn produces a high intensity light source by the FEL. During the operation of ALICE hazardous radiation is created including x-rays, high energy gamma rays and neutron radiation [102], so as a

precaution the accelerator hall has thick concrete walls to block the radiation from leaving the hall.

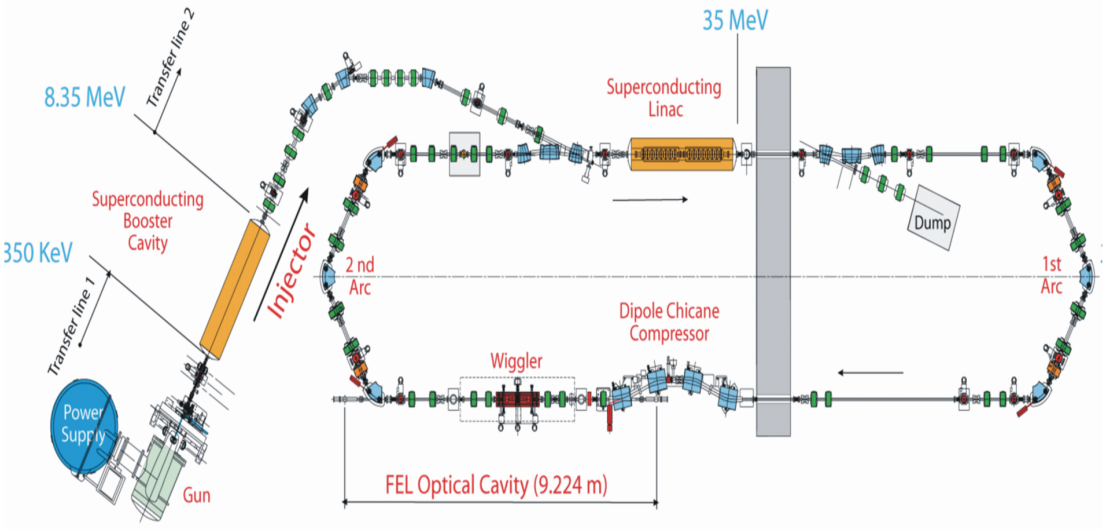

FIGURE 2.10: A detailed schematic of the ALICE accelerator, with the electron bunch energies shown at various position, including the integrated IR FEL. [101]

A high powered laser is first used to free electrons from a donor source. As shown in Figure 2.10, the electrons are then fed into the first of two superconducting linear accelerators which accelerates them from 350 keV to 8.35 MeV. The electron bunch is then accelerated for a second time to 35 MeV by a second linear accelerator. Before entering the FEL the electron bunch needs to be compressed to as small as possible to improve the quality of the IR light it produces. This is done by using a dipole chicane containing four bending magnets which act to compress the electron bunch temporarily before it reaches the FEL.

The FEL works by using a set of powerful alternating dipole magnets called an undulator to force the electron bunch to oscillate as it travels through the cavity between them [103]. This change in direction causes the electron bunch to produce synchrotron radiation. By changing the size of the gap between the two magnets it is possible to tune the FEL radiation to a particular wavelength. The undulator gap was generally varied over 10 to 20 mm which in turn produced light over the range of 5.5-9 µm. By placing the undulator between two cavity mirrors the light produced from the electron bunches is synchronised and therefore produces a coherent IR light beam. Since the light is coherent there will also be

an amplification in the power of the IR beam which is why the FEL is able to produce a very intense IR source.

To properly set up the FEL before a scan a spectrometer was used to check that the FEL was producing light of the desired wavelength and with an acceptable bandwidth. This was an important step because the undulator gap would produce a slightly different wavelength of light day to day due to varying environmental conditions and accelerator settings. During a SNOM scan a feedback mechanism created by Paul Bassan, Andrzej Wolski and the ASTeC group was used to maintain the wavelength of the FEL light by adjusting the gap size to account for any drifts due to changing ambient conditions or instrumental drifts.

The ALICE accelerator was designed to run at 10 Hz, producing an intense macropulse of IR light lasting for around 75 µs. The macropluse is made up of a train of much smaller micropulses running at the rate of 16.25 MHz with each micropulse lasting for 1 ps. A power meter was positioned near the output of the FEL so a measurement of the IR beam's power and stability could be taken before a scan was started. The FEL produced an average power of around 15-20 mW depending on accelerator settings and the wavelength being produced, which given the small duty cycle of around 0.00012% means the IR light peak power is very high. The bandwidth of the FEL light varied between 0.08-0.13 µm at the full width half maximum (FWHM) again depending on the settings used and the wavelength the FEL was producing. The FEL radiation was found to have a short-term stability of ≈1% while running under optimal conditions.

As the FEL is located in the accelerator hall the IR light has to be transferred to the optics table within the diagnostics room. To do this with minimal loses an evacuated beamline, $\approx 10^{-3}$ torr, was used to reduce absorption due to air. The beamline uses large adjustable Al and Ag mirrors to direct the beam along the beamline. They were chosen for their exceptional IR reflectivity of 90% and 94% respectively and also their low cost compared to other options such as Au mirrors. The very end of the beam line is positioned above the SNOM table, as

seen in Figure 2.9, with a CaF_2 window to allow the IR light to leave the beamline efficiently.

2.4.3.2 Optics

As shown in Figure 2.9, many mirrors are needed to direct the IR light from the beamline exit to the sample which is held in the SNOM instrument. The mirrors on the optics table were made of Au with an SiO_2 protective layer and were chosen due to their exceptional reflectance of $> 96\%$ and very uniform optical response over the IR wavelengths used within this study.

The mirrors carry out important secondary role while directing the light towards the SNOM, they also rotate the IR beam by 90°. After characterising the FEL beam, it was found that once it had reached the optics table it had significant spatial structures. With reference to the SNOM instrument, the beam had many irregular features in the vertical axis but had a much smoother structure throughout the horizontal plane. This would be detrimental when imaging in reflection mode since the fibre is fixed within the horizontal plane but varies in the vertical axis as the tip is raised/lowered to follow the undulating sample surface. This could therefore create contrast within the SNOM images which is not due to the variation of the chemicals present but instead due to the tips aperture moving through the spatial variation in the FEL beam. To minimise this effect the first three mirrors the light strikes after leaving the beamline act to rotate the beam by 90°, as shown in Figure 2.11, so that as the tip moves vertically it travels through a smoother axis of the beam. This wasn't a concern for transmission mode since the tip is fixed in the same part of the FEL beam regardless of how the tip moves vertically due to the light being directed to the tip from beneath.

It was also found in previous SNOM runs that there is a small amount of higher harmonics light produced by the FEL. Since this was mostly second order harmonics they consisted of wavelengths between $\approx$ 2.5-4.2 µm. Although the FEL only produces a very small amount of light at higher harmonics it is a serious issue

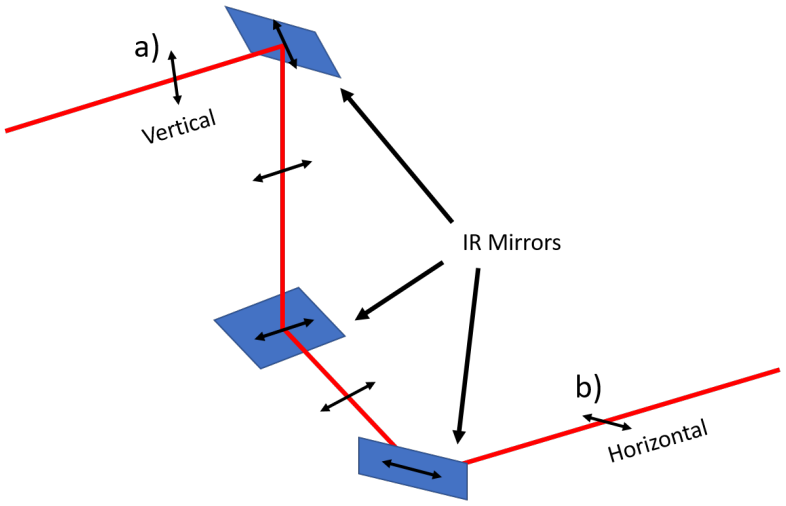

FIGURE 2.11: A schematic showing how the FEL IR beam is rotated by 90° by using three mirrors. As shown the axis is vertical at a) and is horizontal by b).

since the shorter wavelengths are transmitted more efficently within the beamline and the IR fibre used on the SNOM. It could therefore become a considerable component of the signal as seen by the MCT detector, which in some cases was found to be as large as 50%. This could therefore give contrast within the SNOM images due to the absorption at the unwanted higher harmonic wavelengths. This problem once found, was easily overcome by placing a tested high pass filter in the beams path, designed to block IR wavelengths under 4.5 µm. The filter was tested using a single pixel FTIR at Liverpool University and was found to work very well having a transmission rate of $<<$ 1% for wavelengths under 4.5 µm while having excellent transmission over the wavelengths necessary for the SNOM experiments.

A pair of polarisers were used to control the amount of light that reaches the SNOM. By rotating one of the polarisers the amount of IR light which passed through was dependent on the difference in the relative polarisation axes between the two polarisers. The first was housed in a motorised rotating mount while the second was placed in a fixed mount. They were ordered in such a way so that the polarisation of the light which reached the SNOM tip was always constant. Since the power of the FEL light can vary considerably over its operational wavelength range it was an important capability of the SNOM instrumentation to be able to control the amount of light that reaches the sample. The IR-FEL used, produces a very intense light source creating a large amount of light over a relatively small

time period. The power is actually so great that it can cause damage to both the sample and the instrument once the beam has been focused. Since the IR fibres also have a large variation in the transmittance of IR light at different wavelengths, which when coupled with the variations in the power of the FEL light could cause the MCT signal to vary by orders of magnitudes from one wavelength to the next. This is a concern for the detector but also the sensitive electronics used to record the amplified MCT signal. So care was take to block the light completely with the polrsiers each time the wavelength of the FEL was changed, where it would then be slowly rotated to allow more light until the levels reached the optimal point.

A commercially bought CaF_2 window placed at an angle to the IR beam was used as a beam splitter sending $\approx$ 20% of the beam to a pyroelectric detector, which was used as a reference signal, while the rest of the beam was sent to the sample within the SNOM.

The amount of light captured by the SNOM fibre was very small due to the tiny aperture, therefore it was important to get a high density of light underneath the tip to maximise the amount of captured light. To do this commercially bought ZnSe lenses were used to focus the IR light onto a small point directly under the SNOM tip. Since there are two modes of operation and hence two light paths it was easiest to have a lens for each path with an appropriate focal length of 300 mm and 600 mm for reflection and transmission respectively. ZnSe was used instead of other IR transmissive materials, such as CaF_2, since it is able to make sturdy long focal length lenses not practical with other materials. A much shorter focal length CaF_2 lens was used to focus the light onto the pyroelectric reference detector with a focal length of 50 mm.

2.4.3.3 SNOM microscope

The SNOM microscope and its optical components can be simplified into a general layout as shown in Figure 2.12.

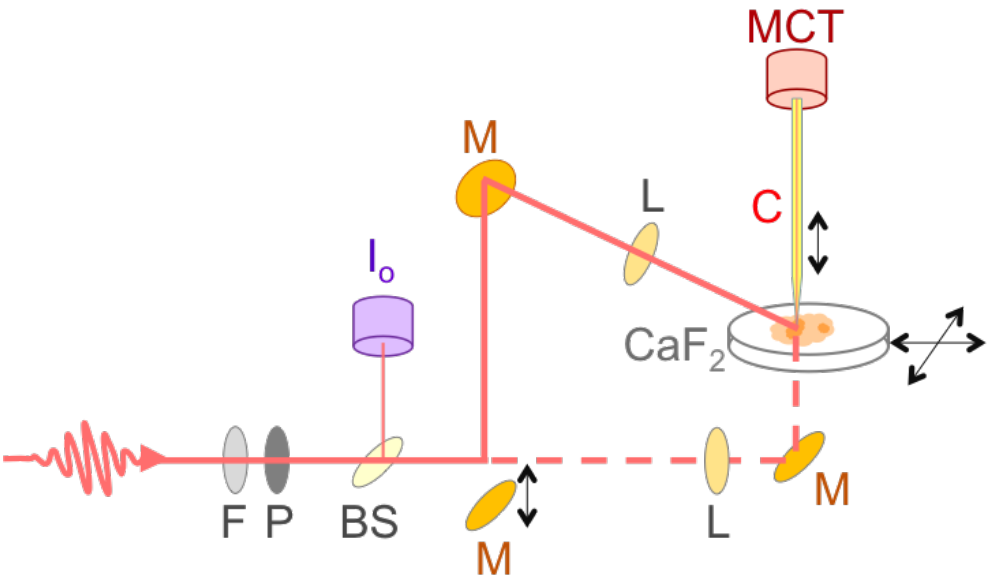

FIGURE 2.12: A schematic of the SNOM microscope with the main components shown. M represent a mirror, L represents a lens, F represents a filter, P represents a polariser, BS represents a beam splitter, C represents the SNOM fibre, I_0 is the reference detector and the MCT is the LN_2 cooled mercury-cadmium-telluride detector.

The base of the SNOM instrument is a model IX3 inverted microscope made by Olympus which has been modified, by various members of the SCAnCan group, for uses in SNOM [104]. These modifications included the addition of a piezo sample stage and SNOM head which contains the components needed to hold and control the IR fibre. A major benefit to having an inverted microscope is that unlike previous iterations of the SNOM it allows for the live imaging of the sample from underneath with an optical camera at various magnifications. This was done using a GXcam optical camera attached to the side of the microscope allowing for the collection of images much like traditional optical microscopes, examples are shown in Figure 2.13. This allows for the sample to be imaged and studied before starting a IR SNOM scan, which helps with finding areas of interest quickly and efficiently. Since the samples used in the SNOM have a maximum thickness of $\approx$ 7 µm, it is possible to see through most biological samples therefore allowing the SNOM tip to be seen. Because of this it is very easy to place the tip directly over any intended area, which can often be a struggle with SPM techniques.

The next major part of the SNOM is the sample stage which is made up of

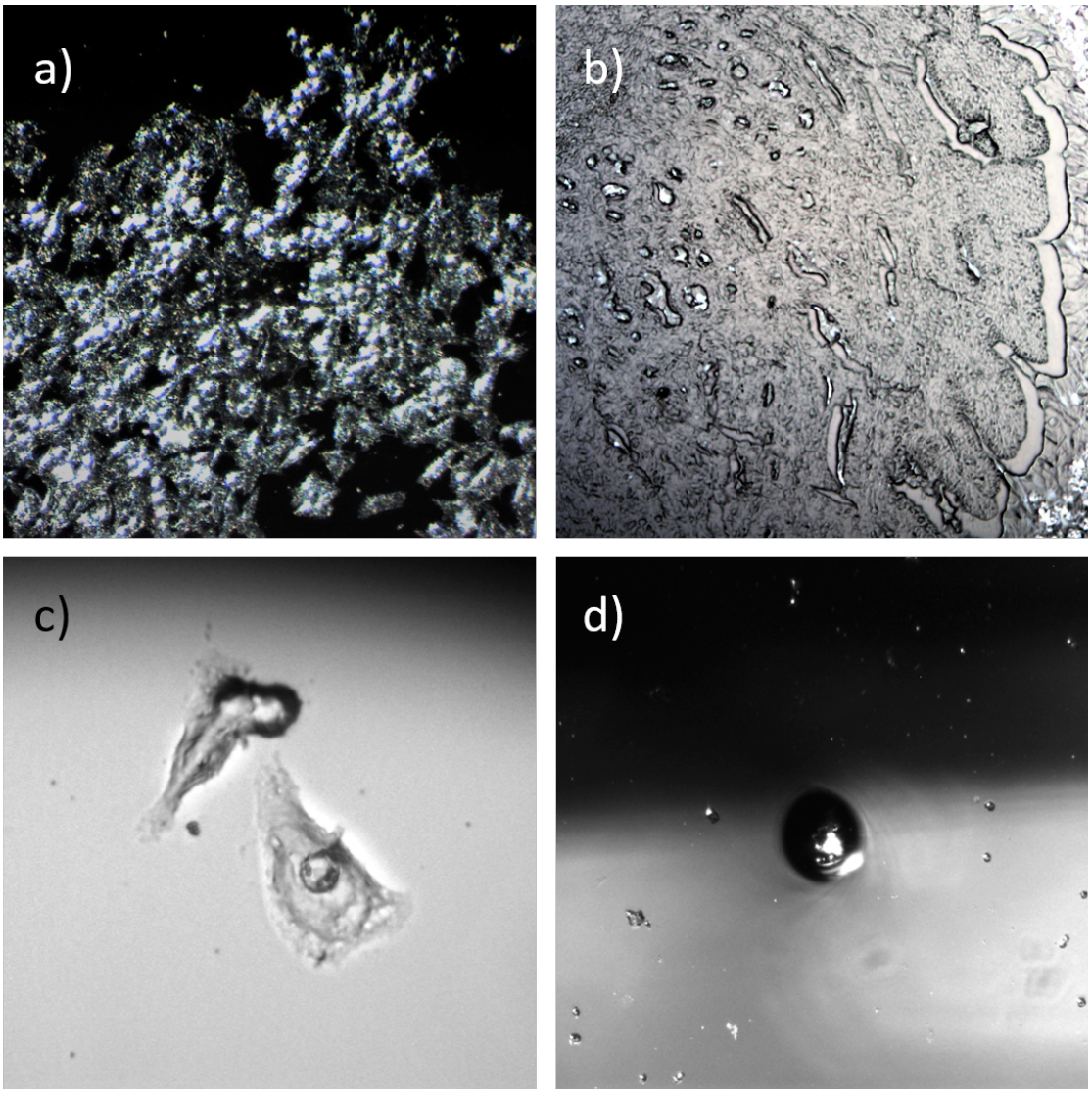

FIGURE 2.13: Examples of images taken by the inverted microscope on the SNOM. a) A low magnification image of a cluster of cells. b) A low magnification image of a tissue sample. c) A higher resolution image of two cells. d) An image where the SNOM fibre is close enough to the sample so that it is in focus.

two components, a large travel manually operated sample stage and a computer controlled piezoelectric stage used for the finer movements needed while scanning. The piezoelectric stage used on the SNOM is a Piezosystem Jena PXY 500 AP dual axis translation stage. It is designed to hold a standard sized sample slide commonly used in traditional microscopy. The coarse manual controls have sufficient travel to allow the tip to be placed anywhere over the sample slide allowing for any area to be imaged. Once the tip is close to the desired area it's position can be checked using the inverted microscope housed within the SNOM base. The

piezoelectric stage can then be used to finely position the tip to precisely image the intended area. The piezostage has a movement range of 500 µm x 500 µm which when compared to the manual stage is very small but gives quite a large scan area compared to most other piezostages. The piezoelectric stage is controlled by the SNOM's software which determines the image size and the number of pixels to be collected. Since the sample stage has a piezostage it is capable of very small pixel sizes in the image therefore allowing for very high detail images to be produced. Unlike with a multi-sensor detector such as the FPA in the FTIR instrument the SNOM is only able to collect one pixel at a time. It therefore has to build the 2D image by rastering the sample underneath the SNOM tip, with each FEL pulse being used to create a single pixel. Since data acquisition rate is therefore coupled to that of the FEL rate (10 Hz) it is important to consider before starting the scan the number of pixels within the image, as the more pixels the longer it will take to complete the scan. The time needed to take a SNOM image becomes the major limitation as it can take close to 1 hour and 15 minutes to collect a 150 x 150 pixel image at a single wavelength.

The final major component is the SNOM head, who's primary role is to hold the fibre and control the vertical movement of the SNOM tip as it moves over the sample surface. A piezoelectric drive is used to raise/lower the tip and to maintain the separation between the tip and sample a set of bimprophs are used to facilitate a feedback mechanism. To attach the fibre to the bimorph it is carefully glued in such away so that only a small amount of the tip $\approx$ 3-5 mm overhangs the bottom of the bimorph. Foam fittings are used within the SNOM head to help support the fibre minimising the stress on both the fibre and glue.

A bimorph is a device made up of two piezoelectric ceramics which sandwich a nonactive layer of titanium, as shown in Figure 2.14. By applying inverse AC voltages to each of the piezoelectric materials within a bimorph it can be made to oscillate from side to side. This is possible since the two materials receive opposite voltages so as one side of the bimorph expands and the other contracts, as demonstrated in Figure 2.14. This type of bimorph is called a driving bimorph

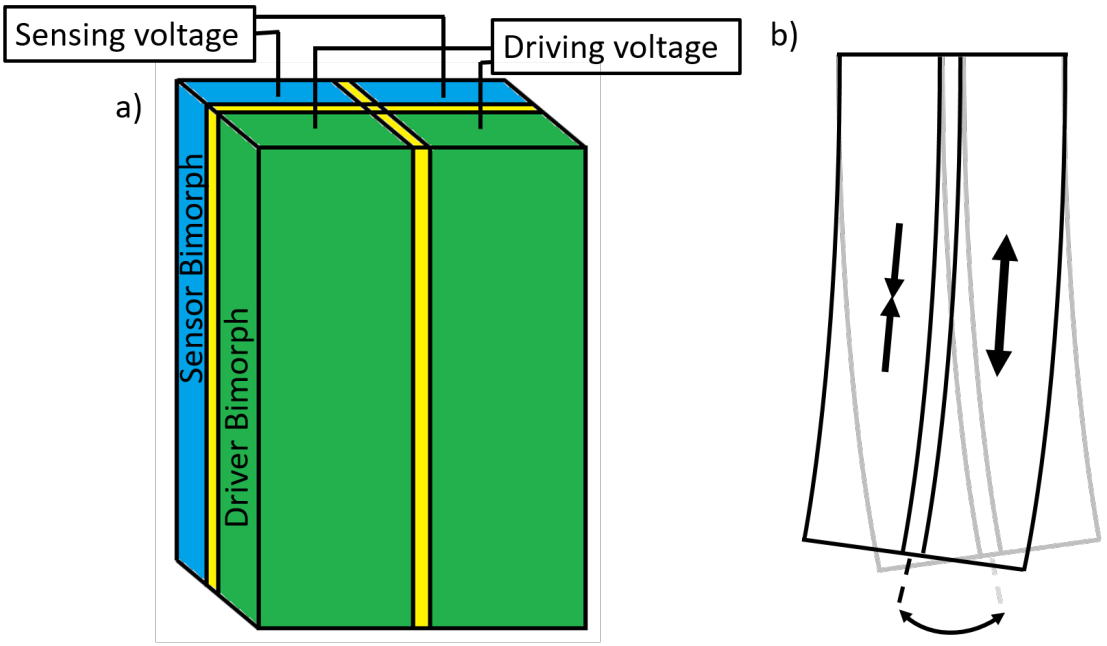

FIGURE 2.14: A diagram showing the biomorph configuration in the SNOM head. a) Shows the driving bimorph (green), sensing bimorph (blue) and the non-piezoelectric filler material (yellow). b) Shows an exaggerated example of how by applying an alternating voltage the end of the bimorph is able to swing.

since it receives a driving voltage which forces it to oscillate. By changing the amplitude of the voltage the size of the oscillation can be adjusted. Since the voltage is what drives the oscillation it is also possible to change the frequency of the oscillation by adjusting the frequency of the driving voltage.

Attached to the driving bimorph is a second bimorph called a sensor bimorph, as shown in Figure 2.14, which is identical apart from it doesn't receive a driving voltage. Just as when a voltage is applied to a piezoelectric material it expands or contracts, the reverse is true meaning that if a piezoelectric material is compressed or stretched by an external force it creates a measurable voltage; this is the basis of how sensor bimorphs work. As the sensor bimorph is forced to oscillate due to the driving bimorph it produces a reference voltage which is monitored by the SNOMs software. The resonant frequency of the bimorphs and the attached fibre can be found by varying the frequency of the driving voltage as to maximise the reference signal, which was done before each scan. As the tip is lowered to the sample it begins to interact via the Van der Waals forces present at the samples surface. This force acts to dampen the oscillation of the tip, which would be seen

as a drop in the reference voltage. Once the tip is at the desired distance from the samples surface the reference signal is recorded and the SNOM can begin imaging. If the sample were to move away from the tip as it was rastered underneath, the reference signal would increase due to less dampening by to the Van der Vaals forces. A piezoelectric drive with a travel of $\approx 10\,\mu m$, to which all the bimorphs are mounted, would then lower the tip until the reference signal returned to the recorded value which indicates that the tip to sample distance as been returned back the the original distance. The opposite works as one would expect, if the sample were to rise the dampening would increase causing the reference signal to decrease. This would cause the SNOM tip to be retracted until the set tip to sample distance was restored. By recording the voltage supplied to the piezo drive responsible for raising/lowering the fibre as it scans over the samples surface, a topographical image is produced since the voltage is proportional to the height of the sample. This allows for both the topography and variation in absorption to be imaged in every scan. An important setting while configuring the SNOM for a scan is the gain, which controls the speed at which the tip responds to the changing height of the surface. For samples with large variations in height a higher gain would be preferable so the tip would respond quicker to the changes. Whereas if the sample is instead very flat, as is the case for a lot of non-biological studies, the gain would be lowered so that the tip is not jittering due to it over-correcting, which could lead to noisier images being produced. This was an important setting to become accustomed to as there was often a 'sweet spot' to give the best images.

2.4.3.4 Optical fibres

As the SCAnCan group's SNOM instrument uses an optical fibre to both create a tiny aperture for near-field imaging [105] and to also efficiently transmit the captured light to a MCT detector. It is therefore critical that the material used for the optical fibre's inner core has a high transmittance for the light being used, which in this case is IR radiation. Visible wavelength optical fibres have been used for some time and are commonly used all around the world for a wide variety of applications, such as telecommunications, industry and research. IR fibres have

only recently become commercially available allowing for the development of IR-SNOM. One of the major difficulties involved in making an appropriate IR fibre for SNOM imaging lies in making an inner core which has a very small diameter while being over to a meter long. The fibre must also be flexible, robust and most importantly efficient at transmitting IR light. For the IR-SNOM experiments shown in this thesis the fibres used were made by CorActive [106] and had a inner core made of selenide glass As_2Se_3. Various sized cores were tested with diameters of 6 µm, 14 µm, 20 µm and 100 µm being studied. The inner core is surrounded by a layer called the cladding, which independent of the inner core's size had a diameter of 170 µm, it is the cladding material which is seen in Figure 2.7. The cladding protects the fragile inner core and also helps the transmission efficiency of the fibre due to total internal reflection discussed later in this chapter. Over the cladding there are more layers, including a acrylate layer, which help shield the core from outside interference and also give the fibre much more strength allowing them to be easily handled and usable on a daily basis.

The most commonly used fibre for SNOM imaging was the 6 µm core. All but the 100 µm core fibres work in single mode operation which means the core's diameter is small enough that the light is forced to travel in a single path. By operating in single mode there is very little dispersion compared to the larger cored fibres which operate in a multimodal operation which allows the light to travel along multiple paths. It is preferable to operate in single mode since the light signal is very small and any dispersion of the light pulse both spatially and temporally would weaken it further. This is a much more important effect when the SNOM is used in conjunction with the lower intensity light sources such as quantum cascade lasers which have a much weaker peak power compared to a FEL.

Independent of the type of light used optical fibres all use the same principle of total internal reflection to allow the light to travel along a curved fibre with very little losses. For the light to be totally internally reflected the cladding which surrounds the inner core of the fibre must have a refractive index which is less than that of the inner core. The light rays which are incident on the boundary between

the core and cladding at an angle (θ) which is greater than the critical angle (θ_c) will be completely reflected back into the inner core. The critical angle is defined by Snell's law which is shown in Equation 2.22, where n_2 is the refractive index of the cladding and n_1 is the refractive index of the inner core.

$$\theta_c = sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{2.22}$$

Snell's law can be used to fine tune the refractive indices of both the inner core and the cladding to give the correct conditions for total internal reflection. Although having a very small critical angle would allow for more light to be reflected and hence less losses, it can lead to greater dispersion in multimodal fibres as there are more paths the light can travel given the larger acceptance angle. The fibre's material are therefore designed to optimise both the transmission efficiency of the fibre and also the dispersion of the signal.

The fibres were also used to create tiny apertures appropriate for near-field imaging. One method would be to cleave the end of the fibre so that the aperture is simply the exposed inner core giving an aperture size equal to the inner core's diameter. This method was used in the earlier studies with cores as small as 6 µm, but this method would not give the submicron resolution desired for the later experiments. To create much smaller apertures the fibre needs to be sharpened and then coated in a opaque metal to create a very small aperture over the fibres inner core. To do this ≈ 2-3 cm of the acrylate protective layer were removed from the end of the fibre using a scalpel so that cladding layer was exposed. To etch the tip into a sharp point a solution of 3 parts hydrogen peroxide (Sigma) and 7 parts sulphuric acid (Fisher Scientific) was used, this mixture is called piranha solution. On top of the piranha solution a thin layer (≈ 2-3 mm) of tetramethylpentadecane (TMPD) solution (Sigma) was added. The fibre was then lowered and held in the piranha solution so that at least 5 mm was submerged within the acid. Figure 2.15 shows how the piranha solution and the TMPD layer are used to form convection currents within the acid which eat away at the fibre near the border between the two solutions until eventually the end of the fibre falls off [107]. The result is

a fibre which has been sharpened to a very small point appropriate for SNOM imaging. Once the fibre was fully etched it was quickly removed and was cleaned with methanol and dried with Argon gas to remove any remaining solution.

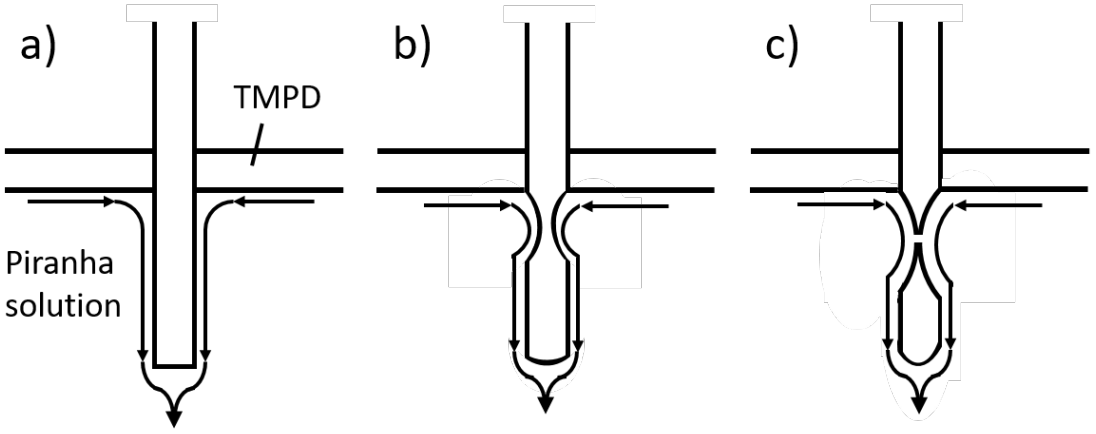

FIGURE 2.15: A diagram showing the stages in tip etching. a) The tip was submerged in piranha solution topped with a protective layer of tetramethylpentadecane (TMPD). b) Due to the convection currents the tip begins to be etched near the boundary of the two soultions. c) The end falls off leaving a sharpened tip on the fibre.

Once the tip has been etched the next step is to coat it in a thin layer ≈ 30 nm of an opaque metal so that the entire tip is evenly covered apart from a tiny aperture over the core. This was done by placing the fibre in a specially designed holder capable of rotating within an evacuated chamber so that the fibre's tip is able to rotate around its axis. A tungsten filament with gold wire wrapped around it was used as a thermal evaporator to evenly coat the fibre's tip as it was rotated. A thin film thickness monitor was used within the chamber to measure the approximate thickness of the gold layer added to the fibre so that a sufficient layer gold layer on the fibre could be assured.

2.4.3.5 Detectors

Within the SNOM instrumentation two detectors were used in unison for every scan, a MCT detector and a pyroelectric detector. The MCT detector was used to measure the light collected by the SNOM fibre, while the pyroelectric detector measures a portion of the FEL beam which has been diverted at the beam splitter

previously discussed. The recorded variations in the FEL beam's intensity create a reference signal which was used post scan to help remove features in the SNOM images due to fluctuations in the FEL beam intensity, this process is outlined in further detail in the normalisation section. Two different types of detectors were chosen because even though the diverted beam is only $\approx 20\%$ of the FEL light that reaches the SNOM it is still many magnitudes greater than the light captured by the SNOM fibre. Appropriate IR detectors with different sensitivities were therefore selected for each of the light sources.

For a low intensity IR signal such as the light captured by the SNOM fibre a single element N_2 (liq) cooled MCT detector is ideal. The MCT detector utilised its tuneable bandgap HgCdTe sensor, hence the name, to measure IR signals with a very high sensitivity. As a IR photon strikes the sensor an electron is excited from the valance band into the conduction band, creating an electron and hole pair. A bias current is applied to across the sensor causes the free electrons and holes to move depending on the electric field leading to a measurable pulse in the voltage. The more photons which strike the sensor the more excitation's will occur and hence a larger voltage will be recorded. Since the sensor is very sensitive in the IR ranges it has to be housed within a vacuum and cooled with N_2 (liq) to minimise thermal noise, which is considerable in the IR range. The end of the IR fibre and the detector window is surrounded by a cover to block any external IR radiation which may interfere with the SNOM measurements. Very weak signals such as the voltage produced by the MCT are prone to gaining a lot of noise as they travel along wires. The MCT's output was therefore amplified as soon as possible after leaving the MCT detector so that it is much larger in magnitude than the noise. The main limitation of the MCT detector is that is only able to operate over relatively low intensity levels. A large IR signal would saturate the detector and may even damage the sensor within. This makes it inappropriate to measure the FEL reference signal. Although the MCT is very sensitive to the intensity of the IR light it isn't time sensitive enough to resolve the individual micropulses which make up the macropulse. This is not a problem though as the train of micropulses were intended to act as one large macropulse.

The second detector used in the SNOM instrument was a less sensitive but with a higher saturation point single element pyroelectric detector capable of measuring the high intensity of the diverted FEL beam. A pyroelectric detector works differently than a MCT detector since it doesn't rely on the creation of electron-hole pairs and instead utilises the pyroelectric properties of crystals which are temperature sensitive. The thermally sensitive crystals within IR detectors use coatings designed to absorb incident IR light, which in turn heats the crystal sensor. In pyroelectric sensors a change in it's temperature causes it's surface charge to change which therefore alters the polarisation of the crystal. The change in polarisation of the sensor enables the detector to give a voltage output which is proportional to the intensity of the incident IR radiation. A relatively large number of photons are required to significantly change the temperature of the crystal sensor, so it is much better suited to measuring the high intensity FEL reference beam, unlike a MCT detector.

Both of the detectors used in the SNOM set up gave a continuous reading for the IR light intensity. But since the FEL generates short pulses of IR radiation which are seen as pulses in the detector outputs, it isn't appropriate to simply average the signal from one pulse to the next as there is a large amount of time when there is no IR light striking the detectors. To account for the pulse structure seen by the detectors both outputs are sent to separate boxcar integrators. The boxcars are used to integrate the detector outputs over a predefined period, called a window, which covers the majority of a pulse seen in the outputs due to a IR FEL pulse. A background reading is also recorded with the boxcar by integrating over a section of the detectors output between pulses where there is no IR light. Each of the windows were set to integrate over the same length of time so that they could be directly compared. The background measurement is taken away from the measurement taken over the pulse and is output from the boxcar as a voltage. This voltage is directly proportional to the intensity of IR light over a defined section of the pulse and it is this value which is recorded by the SNOM software. The windows which determine where the detector signal is integrated are given by an offset from a trigger pulse which runs at the same rate as the FEL

macropulses (10 Hz). Since the FEL pulses are very periodic this means the same portion of every pulse is integrated over giving a reliable measure of the intensity variation for each pulse. By integrating over a range of values in both the pulse and background regions of the detector outputs the higher frequency noise present is greatly suppressed in the boxcar output.

2.5 Sample information

Throughout this thesis experiments are shown using a variety of samples including tissue biopsies and cell lines. A deeper discussion into the reasoning and significance of imaging these samples will be done later in the relevant chapters. Below will outline how the samples were collected, stored and prepared for imaging.

All the samples were mounted onto the same 2 mm thick CaF_2 sample slides (Crystran Ltd, Poole, UK), which were chosen for their relatively cheap cost and high transmittance of IR light, critical for both the FTIR and the transmission SNOM imaging.

2.5.1 Barrett's oesophagus tissue samples

Oesophagus tissue was collected via biopsies at the Royal Liverpool and Broadgreen University Hospitals from patients who had been diagnosed with either Barrett's oesophagus with no dysplasia present in their histology results, or Barrett's associated oesophageal adenocarcinoma. All patients gave their consent for the samples to be taken and were over the age of 18. Once the tissue had been removed it was fixed using 10% formalin and embedded in paraffin wax which allows the sample to be safely stored with minimal degradation. To produce the thin samples needed for FTIR and transmission SNOM imaging the tissue still embedded in wax had to be sliced into 5 µm thick sections by a microtome. By placing the tissue slice in water where they would float it is possible to carefully position the sample on the CaF_2 slide. The samples were then stored in a clean slide holder to prevent

contaminants from settling on the slide. Figure 2.16 shows a typical example of Barrett's oesophagus tissue which is embedded in wax alongside a adjacent slice from the same sample which has had the wax removed and stained with H&E. It is clear that the H & E stain gives a lot more information on the chemical structure of the sample when using microscopy which is why it has become the world standard.

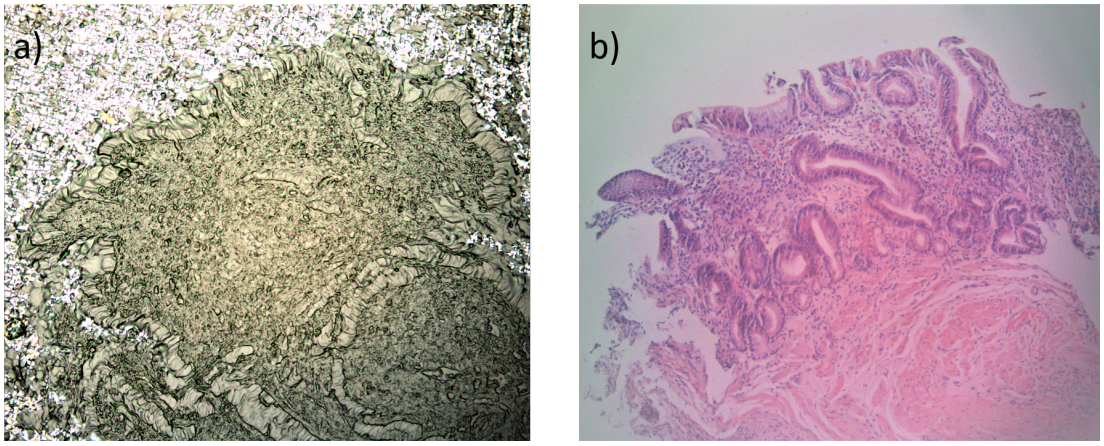

FIGURE 2.16: An image comparing a two different slices from a Barrett's oesophagus tissue sample; a) is an image of a unstained slice which is still embedded within wax, b) shows an adjacent slice which has been dewaxed and stained with H & E.

For FTIR imaging the tissue samples can be imaged with the wax still present so there is no preparation needed before imaging. There will be features due to the wax within the IR spectra but since they aren't in the key biological areas and are well known they are easily removed post scan.

For SNOM imaging the wax must be removed as it is not fixed in place and could coat or even damage the SNOM tip. Xylene (Sigma) was used to dewax the sample slides and it also acts to remove most other contaminants which may have settled on the samples surface. To ensure that all the wax is properly removed the slide is submerged in three successive baths of xylene for 15-20 minutes each time. Once the slide has been fully dewaxed it was rinsed in isopropyl alcohol (Sigma) and left to dry for around 2-3 hours before imaging was started.

2.5.2 Barrett's cell lines

As well as imaging tissue samples which can be very complex systems with a variety of different tissue types within it, cell lines were also imaged where each slide only contained cells of one given type. In total five different types of cell lines were grown with each being given a label, table 2.1 defines what each cell line is.

The OE 19, OE 21 and OE 33 samples where obtained from HPA Culture Collections (Sigma, Dorset, UK). The primary myofibroblast samples were collected from consenting patients who had undergone surgery for oesophageal cancer. As shown in the table OE 19 and OE 33 are the same type of cell type but OE 19 is taken from the oesophageal gastric junction and OE 33 is taken from the lower oesophagus. The myofibroblast samples were taken from the same patient who had a Barrett's background. The 173/1 sample is described as cancer associated myofibroblasts, which means that they were taken from an area which had been diagnosed as being cancerous. The 173/5 sample was taken from an area of healthy tissue which bordered the cancer.

Cell label	Cell type
OE 19	Human Caucasian oesophageal adenocarcinoma (oesophageal gastric junction)
OE 21	Human Caucasian oesophageal squamous cell carcinoma
OE 33	Human Caucasian oesophageal adenocarcinoma (lower oesophagus)
173/1	Cancer associated myofibroblast
173/5	Non cancer associated myofibroblast

TABLE 2.1: A table describing the type of cell associated with each cell type label.

The cell cultures were grown on CaF_2 slides which had been cleaned using ethanol, then rinsed with ultra pure water (18.2 MΩ/cm) (Millipore, Watford, UK) and finally irradiated with UV light for 30 minutes to ensure they were properly sterile which is critical for cell line studies. The cells were seeded on each slide and incubated in an environment of Roswell Park Memorial Institute (RPMI 1640) growth media (Sigma) supplemented with 2 Mm glutamine (Sigma), 10% v/v

fetal bovine serum (Invitrogen, Paisley, UK) and 1% v/v penicillin/streptomycin (Sigma) at 37° C in a 5% CO_2 atmosphere. After around 2 days the cells covered $\approx$ 60-70% of the slide at which point the growth media was removed and the cells were fixed using a 4% v/v paraformaldehyde (PFA) solution (ThermoFisher Scientific, Loughborough, UK).

The cell lines were then stored at 4°C in a phosphate buffer solution (PBS) until the samples were needed for imaging. Then approximately 90 minutes before imaging the desired cell line slide was removed from the PBS and carefully rinsed using ultrapure water (18.2 MΩ/cm) (Millipore, Watford, UK) multiple times to ensure all the PBS solution has been properly rinsed off. The PBS has to be removed so that it does not interfere with the collected IR spectra. After being rinsed the slide has left to air dry before imaging commenced.

Chapter 3

Metric Analysis

3.1 Machine learning assessment

To explain why the Metric analysis was developed for these biological studies, a general understanding of ML techniques is needed. The ideal ML technique has to meet the needs of both clinicians and researchers. A clinician needs a classifier, capable of accurately predicting the type of tissues present within the sample. In cancer research a 'blackbox' approach which is able to accurately predict the tissue types but gives no clear indication as to why the samples were different. As this doesn't provide any insight as to the mechanisms at play in cancer. An ideal ML algorithm would therefore be capable of both.

Work with a type of unsupervised classification algorithm called cluster analysis (CA) was carried out by Timothy Craig [74] on the same type of samples used in the tissue biopsy study. Unsupervised means that the CA algorithm does not use previously defined spectra to learn from, unlike supervised ML techniques which use pre-labelled spectra to train a classifier model. For the sake of brevity only an overview of the cluster analysis work will be discussed here as it is documented in detailed in full within Timothy Craig's thesis [74].

Metric analysis aims to group the spectra into clusters so the spectra within a cluster are more similar than the spectra belonging to another cluster. A 2D

example of this is shown in Figure 1.2, but in principle this can be done in any number of dimensions were each dimension is a variable. As using more dimensions will dramatically increase computational time it is often important to limit the number of dimensions. This is why CA is often used with principle component analysis (PCA), which is able to minimise that size of the datasets while still retaining the key discriminatory variables. As the FTIR data used within this study have close to 1500 individual wavenumbers, this would be to many to feasibly process with CA, so the wavenumbers are restricted to the fingerprint region of 1000-1800 $cm^{-}1$ and PCA is further used to lower the size of the data. Note that PCA is not used within MA.

To group the spectra they are all plotted in a n dimensional graph where n is the number of data points still used after the data minimisation. Similar spectra will be located closer together in the n dimensional space. Multiple centroids are randomly placed within the data space and the distance from each centroid to every the spectra is calculated. The spectra are then grouped based on which centroid they are closer to. The centroid is then moved to centre of all the spectra which were assigned to it's group. As the centroid has moved, the distances to all the spectra needs to be recalculated, at which point the spectra are regrouped based on which centroid they are now closest to. This step is continued until the centroids no longer move, at which point the spectra are finally grouped based on their proximity to the centroids. Once all the spectra have been assigned to a cluster, a 'labelled' image can be created which displays spatially areas of the sample belonged to which cluster.

A popular variant of CA is called k-means, which uses a number of centroids defined by the user. The advantages of k-means is that it is simple and often particularly quick when combined with PCA. The downsides of k-means is that there is often high variation within the clustering of the same data as the results are sensitive to the starting locations of the centroids. Clusters can also become 'lost', which means they get stuck to a single anomalous spectra. The results of k-means are also very sensitive to the number of centroid used.

K-means was studied by Timothy Craig on tissue samples of Barrett's oesophagus and oesophageal adenocarcinoma. Craig found that if the correct number of centroids was chosen the clustered spectra would mirror the expected distribution of the various tissue types in the sample. This implies that the k-means was a successful ML technique for tissue classification. Craig concluded though that the methodology was not appropriate for tissue sample discrimination as the user would have to often vary the number of centroids repeatedly to find the optimal number which best represented the data. This would be because clusters would be taken up by contaminants within the FTIR image or anomalous spectra, who's spectra would appear dramatically different than the tissue sample. The issue is that this technique reintroduces the user bias back into the classification process. He also notes that a fundamental flaw with k-means is that even if it classifies the spectra well it gives the user no indication as to why the spectra were grouped, hence this is not appropriate for the researchers needs.

The lack of control over what the CA models clustered to, actually highlighted that unsupervised techniques are not what is needed for tissue discrimination. Ideally a model could be trained using previously labelled spectra so that in the future it could predict which sample type a spectrum of unknown origin was most likely to be. An example of a supervised ML algorithm is random forest (RF). Random forest works by producing multiple decision trees which in turn contain multiple decision nodes which can be optimised by the pre-labelled spectra. Spectra which vary significantly in their structure will take different paths along the decision trees, resulting in different end points. The path of a spectrum with a unknown origin can be compared to those characterised in the learning process. As RF is well established and may meet the needs for tissue discrimination it will be compared to MA. RF is known though for over fitting in the learning process which means that the model is often not flexible [108].

Neural networks is another variant of supervised ML which was considered. They have been shown to perform well as a 'blackbox' classifiers [109, 110] but they are notoriously difficult to interrupt. This means that no indication as two

why the samples were distinguished could be made and therefore did not met the aims needed for the ideal classifier.

As many of the ML techniques which were either tested or considered were found to be inappropriate or not capable a meeting all the needs of the defined ideal classifier, it was decided to develop a new methodology which was specifically designed to fulfil the desired aims. MA is a supervised ML techniques which is robustly able to characterise important features within the learning process, but critically can do so in a way that allows for clear and intuitive insights into why the samples were different. The rest of this chapter will outline the processes at work within the MA algorithm using simplified artificial data for clarity.

3.2 Metric analysis explanation

The main aim of this chapter is to explain how the metrics analysis (MA) code works, by outlining each step in detail and visually demonstrating how the data is processed. The MA algorithm can be complicated and sometimes convoluted due to the many steps and various key terms needing to be understood. For the sake of clarity this chapter will use artificially generated data to help explain the processes within the MA code. As real biological IR spectra are very complicated with many features, the data used within this chapter is overly simplified to help facilitate the explanation. The MA code can be broken down into three main stages; training, testing and analysis stage, the description of each will be outlined in a separate section for clarity. Figure 3.1 is a simplified flow chart which highlights how each of the stages are connected and how the data is processed.

Training stage

Before proceeding to the explanation, it is first critical to understand how the data is structured when it is supplied to the MA code. As MA is a supervised ML method, the code learns to classify the data by studying pre-labelled data with

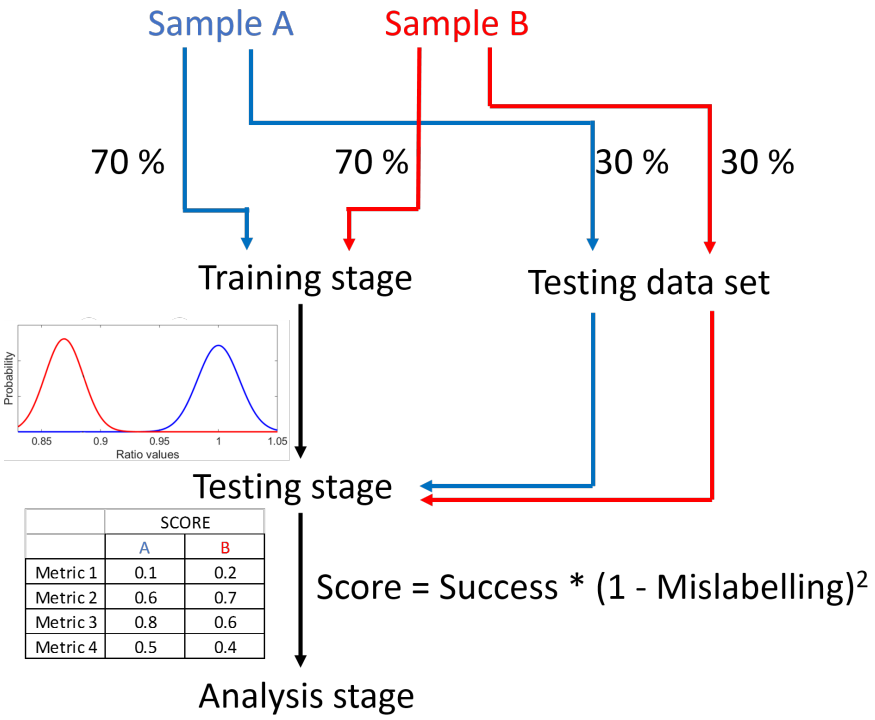

FIGURE 3.1: A flowchart demonstrating the key stages within Metric analysis and how the inputted spectra for both samples are split between the testing and training stages.

known origins, which for this example will consist of two types (sample A and B). The data inputted to MA is therefore l packets, where l is the number of samples and each packet contains spectra representative of only a single sample. How these packets are generated for biological samples varies depending on the sample type and will be explained in the following chapter. Once the data has entered the MA code each packet is split randomly into two, forming a training set (70% of the original packet) and a testing set (30% of the original packet) as shown in Figure 3.1, these datasets are used for different purposes which are outlined further in the MA explanation. For the example shown within this chapter each packet contained 1000 spectra, so 700 are used for the training stage and 300 are saved for the testing stage.

The training stage is where the algorithm attempts to characterise the features of the IR spectra within the training dataset. For clarity consider a single simple spectrum such as the example shown in Figure 3.2, a spectrum which contains two Gaussian peaks. The first concept to understand within MA is that it doesn't

focus on the absorption values at single wavenumbers, it instead analyses the ratio of the absorption values at two wavenumbers.

Doing so has many advantages, one benefit is that a ratio value adds a second later of information to a single data point as it describes the relationship of absorption between two wavenumbers. As has already been discussed absorption at given wavenumbers correspond to specific molecular bonds and therefore specific molecular species. By taking a ratio it is therefore possible to study the relative relationship between biological molecular species within the samples, for example the relative relationship between DNA and glycoprotein which has been shown to be important in previous studies [69].

The second reason for ratios being advantageous is that as the Beers-Lambert law states, absorption is linearly dependent on the thickness of the sample, this results in a ratio value being thickness independent as the thickness components will cancel. If ratios are not taken the spectra are thickness dependent and if studied with no prior processing, differences within the spectral features may occur due to both the variation in the chemical composition and the variations in sample thickness. This would have to be accounted for by some form of normalisation, such as normalising the area under the spectra or equating all spectra to have the same value at a given wavenumber. These methods use assumptions which may not be strictly true, such as all spectra having the same total absorption, but by using ratios the problem is completely avoided.

Evidence of how characteristic information of a spectrum can be inferred from the ratio values can be demonstrated by studying Figure 3.2. For example, if the ratio of the absorption at $\bar{\nu}_1$ (1600 cm^{-1}) is divided by the absorption at $\bar{\nu}_2$ (1560 cm^{-1}), the result would be $0.8/0.5 = 1.6$. This means that $\bar{\nu}_1$ has 60% more absorption than $\bar{\nu}_2$, and hence the ratio is inferring a given relationship between the absorption at two wavenumbers. Obviously depending on the pair of wavenumbers chosen the ratio value will vary such as $\bar{\nu}_1$ divided by $\bar{\nu}_3$ (1400 cm^{-1}) produces a ratio value of $0.8/0.2 = 4$ and $\bar{\nu}_2$ divided by $\bar{\nu}_1$ giving $0.5/0.8 = 0.625$. The key point here is that by taking ratio values it is possible to indicate the relative

relationship between two wavenumbers. In the future the term 'ratio pair' will denote the pair of wavenumbers which are used to produce a given ratio value, ie the ratio pair of $\bar{\nu}_1$ and $\bar{\nu}_2$ produce a ratio value of 1.6. The first wavenumber stated in the ratio pair will always denote the numerator with the second defining the denominator of the ratio.

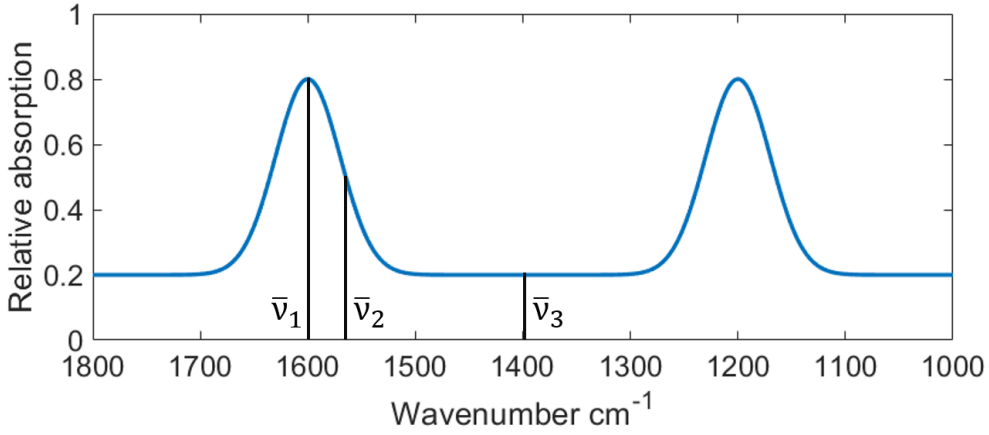

FIGURE 3.2: An example of a simple artificial spectra.

Although studying a ratio pair from a single spectrum may give some insight, in general studies of IR data do not focus on a single spectrum for a given sample but instead use thousands. This is so the learning process is robust and is able to gauge the normal degree of variation between the spectra, which is due to noise and chemical variation within the sample. This is important as the variance at particular wavenumbers will often vary, meaning some wavenumbers may be more reliable as a sample discriminator than others. It is therefore important not to study a single ratio value from a single spectrum but instead the distribution of many ratio values.

To demonstrate this many spectra such as the example shown in Figure 3.2 were generated but with each having a small amount of randomly generated noise added. The absorption at a given wavenumber will therefore have a degree of variation between all the spectra. Figure 3.3 shows the distribution of the ratio values of $\bar{\nu}_1$ divided by $\bar{\nu}_2$, as defined in Figure 3.2, as a histogram. The Gaussian like distribution is also commonly seen in the ratio distributions of biological IR spectra. The greater the variation in the ratio values the wider the histogram

produced, and therefore the histogram is able to describe both the average ratio value along with an indication as to the variance of the ratio values for a given ratio pair.

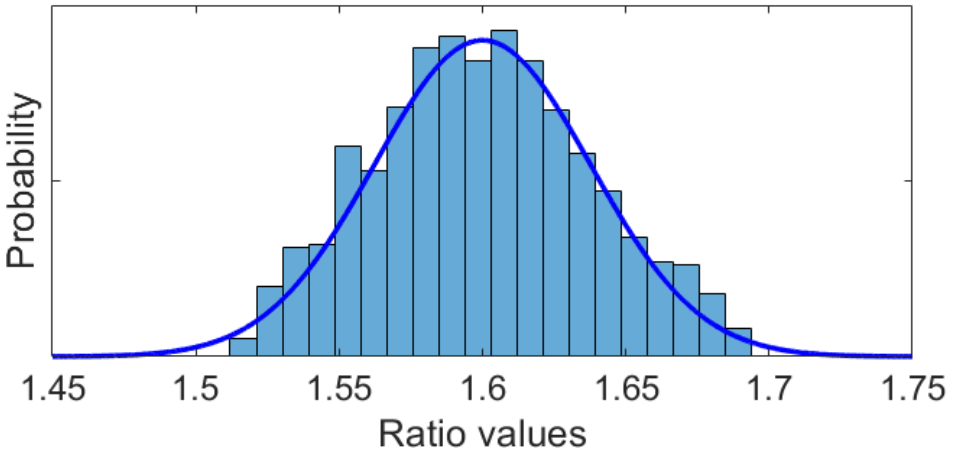

FIGURE 3.3: An example of the distribution ratio values generated by many similar spectra for the ratio pair of $\bar{\nu}_1$ (1600 cm^{-1}) and $\bar{\nu}_2$ (1560 cm^{-1}) as defined in Figure 3.2. The line plot defines the probability density function which describes the ratio values distribution.

A Gaussian probability density function (PDF) which describes the likely probability of a specific ratio value occurring for a given ratio pair can be generated using the distribution of ratio values. The PDF for the data shown in Figure 3.3 is displayed on the same figure but has been scaled so it is clearly visible for the sake of clarity. The PDF is generated using the '*fitdist*' function within MATLAB. The fundamental idea of the training stage is therefore that the relationship between two wavenumbers can be characterised as a PDF, which is capable of describing the probability of a ratio value occurring for a given ratio pair for a given sample. A PDF is therefore able to predict how a spectra from a particular sample type will behave when a ratio between two wavenumbers is taken.

As MA characterises every ratio pair available rather than predefining some wavenumbers as being important, many PDFs are generated. For the standard spectral range of 1000-1800 cm^{-1} and a resolution of $\approx$ 6 cm^{-1} over 17,500 PDFs are generated for every sample type. This approach may take longer to compute than a minimising algorithm which randomly tests metrics and stops when it finds

a 'good' discriminator, but as the MA only takes minutes to run gaining insight into all the wavenumbers and therefore all the molecular species is preferred.

Testing stage

The testing stage aims to use the PDFs generated in the previous stage to produce an optimised model capable of discriminating the sample types. Figure 3.4 demonstrates how PDFs can be used to discriminate samples.

In Figure 3.4 *i*) two spectra are shown, individually labelled A and B with each spectrum being taken from a separate sample packet. As discussed in the training stage each packet contains many spectra which share the same gross structure but have a small amount of random noise added so that there is a degree of variation among them. There is a clear significant difference between the spectra of sample A and B as the peak centred at 1600 cm^{-1} is present in A and not in B. This is an obvious difference which has been exaggerated to highlight the principles at play within the MA. A difference as obvious as this would not occur in the spectra of biological samples as they tend to appear largely similar even between different samples, hence the need for ML.

Figure 3.4 *ii*) shows the histograms and the associated PDFs for the ratio pair of $\bar{\nu}_1$ (1600 cm^{-1}) and $\bar{\nu}_2$ (1200 cm^{-1}) for both samples A and B as defined in *i*), note the colours are consistent for each sample throughout the whole figure. For sample A the absorption values at $\bar{\nu}_1$ and $\bar{\nu}_2$ are very similar, it is obvious therefore why the distribution shown in *ii*) is centred around 1. It is also clear as to why the distribution for the same ratio pair is shifted for sample B, as there is no peak at $\bar{\nu}_1$ and therefore the absorption value for $\bar{\nu}_2$ is larger than $\bar{\nu}_1$ resulting in a ratio value lower than 1. The separated histograms and associated PDFs demonstrate that by characterising the training datasets with PDFs of given ratio pair, both samples can be distinguished from one another as the structures within the spectra are reliably different.

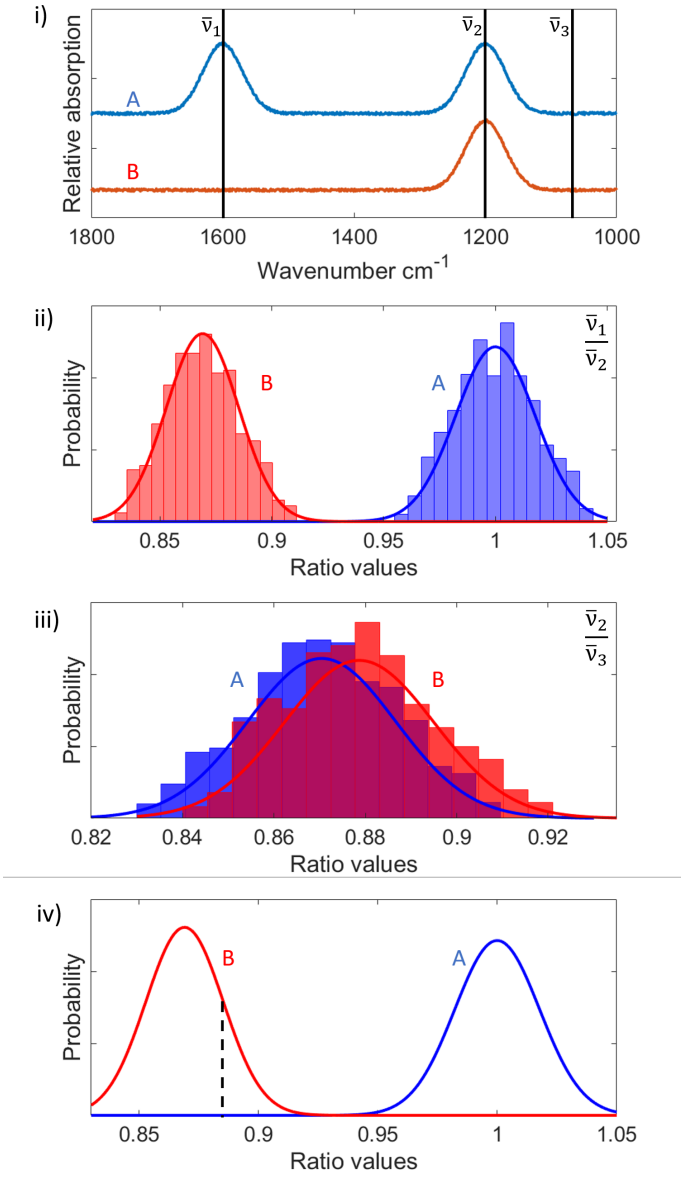

FIGURE 3.4: A demonstration as to how PDFs can be used to discriminate between samples. Note that the colour is used to distinguish the samples with sample A denoted by blue and sample B by red. i) Displays an example spectra for both A and B. ii) Demonstrates an example of a 'good' metric, which uses the ratio pair of $\bar{\nu}_1$ and $\bar{\nu}_2$ defined in i). iii) Is an example of a 'bad' metric, which is made with the ratio pair of $\bar{\nu}_2$ and $\bar{\nu}_3$. (The histogram and PDF for sample B has been slightly shifted to the right to show the relevant histogram and PDF for A, this was done for clarity). iv) Demonstrates how the probability of a given ratio belonging to either A or B can be found.

Figure 3.4 *iii*) demonstrates that not all ratio pairs give significant discrimination, for example the pair of $\bar{\nu}_2$ (1200 cm^{-1}) and $\bar{\nu}_3$ (1080 cm^{-1}). Note that for the sake of clarity, the PDF and associated histogram of B have been slightly shifted to the right to clearly show the presence of the PDF of sample A. In reality the

two PDFs are very similar and therefore are not able to give any discrimination between the samples. If a ratio pair is taken from areas where there is little difference in the relative structures between the two samples the resultant PDFs will be very similar and result in poor discrimination.

An important term to define at this stage is what a 'metric' is, as it is the core idea within MA, hence its inclusion in the name. A metric is a handle term for the multiple PDFs which relate to a single ratio pair, as each sample will have an individual PDF. A metric therefore is capable of describing how a variety of samples are likely to behave when a ratio value between two wavenumbers is taken. The performance of a metric is based on how well it can discriminate a sample from the others. Figure 3.4 *ii*) and Figure 3.4 *iii*) are therefore examples of a 'good' and 'bad' metric respectively. A bad metric is therefore not poorly characterising the data or implying the PDFs are incorrect, it is instead inferring that the PDFs are very similar and are therefore poor at discriminating the samples.

As a metric can describe the likely behaviour of various samples, the spectra within the testing dataset, can be used to asses the performance of each metric. A metric can be used to label a spectra by calculating the probability of the appropriate ratio value occurring for each PDF within the metric. The spectrum is labelled as the sample type associated to the PDF which displays the highest probability. For example, Figure 3.4 *iv*) shows how for a ratio value of ≈ 0.885 the probability of it occurring for both A and B can be determined. Clearly in this example the probability is much higher for B than A, and therefore the spectrum would be labelled as B. All the spectra within the testing dataset are labelled like this and these labels can be compared to the known sample origins to determine the performance of each metric for discrimination.

An ideal metric is therefore one which is able to label all the spectra correctly with no spectra being mislabelled. The performance of a metric can be calculated using Equation 3.1. It is important to note that a single metric gains a score for each sample, as each score only describes the metric's ability to discriminate

an individual sample type. Because the score is sample specific it is described as $score_x$, where in this example x could be either A or B.

$$score_{\mathrm{x}} = successrate_{\mathrm{x}} * (1 - mislabellingrate_{\mathrm{x}})^2 \tag{3.1}$$

The success rate is the rate at which the spectra in the testing set from sample x are correctly labelled as x by the metric. The mislabelling rate is the rate at which spectra which don't originate from sample x are incorrectly labelled x. The success term is an obvious one to include as a 'good' metric should clearly succeed in labelling the spectra of sample x as x, but the mislabelling term is just as critical. If only the success rate was considered a metric which labelled every spectrum a single sample type would gain a perfect score for that sample type. Clearly though this is a poor sample discriminator as it is blindly labelling every spectrum one type regardless it's true origin. The mislabelling rate is therefore key to prevent this as if this were to occur the mislabelling rate would also be large resulting in a poor score for the metric. The mislabelling rate was in fact found to be more important hence why the term is biased by squaring it, which increases it's influence on the score. A good score therefore indicates that the metric is accomplished at correctly labelling the spectra of a given sample while also not incorrectly labelling spectra from a different origins as that sample.

A score is therefore calculated for each sample for every metric, as a metric which is good for one sample is not necessarily good for another if there are more than two samples. Every metric can then be ranked for an individual sample based on their score, with the best ranked having the highest score.

As it would be very hard for a person to evaluate the scores of over 17,500 metrics as a list, a visual method called a butterfly plot was developed to display the score of all the metrics for a given sample within a single image. Figure 3.5 is an example of a butterfly plot generated by the MA algorithm using the simple data shown within this explanation. The axes of a butterfly plot denote the wavenumbers involved within the ratio pair of a metric, with the x axis being

the first term in the ratio pair and the y axis being the second. The pixel colour at the intersection of two wavenumbers therefore denotes the score of the metric which used that ratio pair. For example, 1200 cm^{-1} on the x axis and 1600 cm^{-1} on the y axis produce a very good score, while a ratio pair of 1200 cm^{-1} and 1400 cm^{-1} generates a poor score. Figure 3.5 clearly demonstrates that a metric which contains a wavenumber around 1600 cm^{-1} achieves a very good score as there is a clear difference between the spectra of sample A and B in this region, as shown in Figure 3.2.

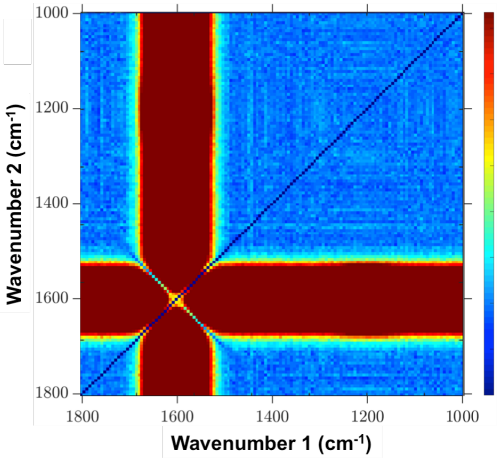

FIGURE 3.5: An example of a butterfly plot, which displays the scores of every metric. A large value (dark red) denotes a good performance and dark blue representing a bad score. Therefore clearly metrics which use a wavenumber around 1600 cm^{-1} perform well which is to be expected within this example.

Analysis stage

The final stage is called the analysis stage, where the best performing metrics are further analysed. If there is a very significant difference between all the samples as is the case in this artificial exaggerated example, then only one metric is needed to distinguish the spectra from sample A and B. Differences of this magnitude do not occur in the biological studies as the variations between the spectra are much more subtle. A single metric may therefore not be able to function as a perfect model and will incorrectly label spectra. This can be combated by using multiple metrics to 'vote' on the correct label for unknown spectra.

Using multiple metrics is beneficial as metrics which use similar ratio pairs which are describing the same spectral features, will reduce the detrimental effect of noise within the spectra. Another advantage is that by having metrics which are studying entirely different regions of the spectra extra information is gained that is useful for discrimination as spectral variations aren't expected to only occur at one feature. By testing the unknown spectrum with multiple metrics a more successful discrimination model is generated. The number of metrics has to be limited to save computational time but importantly to also stop poorly performing metrics having a detrimental effect on the discrimination model. The final stage of the MA algorithm is therefore to asses the highest ranked metrics for each sample, to find the optimal number of metrics which produce a model best capable of discriminating the samples.

This is done by relabelling the testing dataset as was done in the testing stage, but this time taking the label prediction from many metrics rather than just one. The top 300 ranked metrics for a given sample calculate the probability of each spectrum belonging to a particular sample. Each metric then votes for a label based on which had the highest probability, as demonstrated in Figure 3.4 *iv*). Each vote is then weighted based on how large this probability is compared to the probabilities for the other samples. For example, if metric (M1) gave sample A and B a probability of 0 and 1 respectively, M1 will contribute a full vote to B. Another metric (M2) which has a probability for A and B of 0.3 and 0.7 respectively, will also vote for B as it had the highest probability but this vote will have a lower weighting than M1 as the magnitude was lower.

To determine the weighting of a vote all the probabilities are normalised so their sum is equal to 1. The vote is then assigned a weight equal to the largest probability, so the largest vote possible is 1. The result of weighting the votes is that metrics where the largest probability is considerably larger than the others has a greater influence on the final labelling. For example if M1 and M2 were to vote, M1 would give a vote of 1 to B and M2 would contribute a vote of 0.7 to B also. As only the highest probability within a single metric contributes to the vote, M2 wouldn't contribute a vote of 0.3 for A. The votes are then summed for

each sample, which in this example would give 0 and 1.7 for A and B respectively, the spectrum is then labelled as the sample with the highest of these values. This process is carried out while varying the number of metrics used from just a single metric up to 300. This is done for each sample type as each sample will have an individual set of highest ranked metrics. Because of this each sample may also need a different number of metrics to optimally classify the unknown spectra.

To find the optimal number of metrics the labels given by the MA code are compared to the known origins of the spectra and a success rate is calculated. The number of metrics used in the final model is then restricted to the amount which gave the highest success rate. The mislabelling rate is not important here as these highly ranked metrics have already been credited to perform well for a given sample type.

The MA code then outputs an optimised model which can be used to predict the sample type of a spectrum that it has never processed before. This model includes the metrics which have been determined to be the best at sample discrimination. This is all that is required by the clinicians for the 'blackbox' sample classifier. The MA code doesn't just output the model though as all the relevant information needed to analyse chemical variance in detail is also given. As this can often be hard to interpret due to the large amount of numbers the MA also produces many informative figures, including the butterfly plot previously shown, the success curves generated in the analysis stage and the discrimination plots. Due to the simple data used in this example both the success curves and the discrimination plots don't behave similarly to the biological studies and so will be explained in the next chapter.

Chapter 4

FTIR study

4.1 Introduction

As the MA algorithm has already been outlined in detail previously, this chapter aims to demonstrate how it can be applied to biological samples, ascertain how successfully it performs and compare it to a similar widely used classification algorithm.

As discussed in the introduction, the current gold standard method for disease detection and tissue differentiation is to use an optical microscope with various dyes such as H&E, to stain the tissue biopsies various colours based on it's chemical content. A highly trained histopathologist can use the microscope images to diagnose the sample based on its structure and the information provided by the dyes.

FTIR which is the focus of this chapter has emerged as a promising technique for biological studies and has been widely applied to a range of samples. It benefits from being sensitive to the wealth of chemical information contained within IR spectroscopy.

This chapter will present the results of applying the MA code to FTIR data of both tissue biopsies and various commercial cell lines. Previously, examples

of ML techniques called cluster analysis had been applied to FTIR data taken of oesophageal tissue samples by Timothy Craig [74]. The results of which were discussed in the previous chapter and concluded that there is a need for a supervised method that could meet the needs of both the clinicians and potential researchers.

4.2 Data preparation

The FTIR data, as with many experimental techniques needs to be corrected and pre-processed before it can be analysed using MA or any other ML algorithm. The pre-process methods described below act to improve the spectra and also prepare it for analysis.

4.2.1 Background subtraction

As previously mentioned a background image of an area of clean slide is acquired each time a sample is placed in the FTIR spectrometer. By subtracting the background image, features present in the spectra due to absorption in the air by CO_2 and H_2O are removed. A background subtraction also compensates for the varying efficiencies of the pixels within the FTIR's FPA. Accounting for the background is essential when comparing spectra both from within the same image and from different scans.

4.2.2 Mie scattering correction

The interaction of light with an object often results in various scattering effects. An example of scattering is Rayleigh scattering which occurs when the wavelength of the light is considerably larger than the object it is striking, and results in the elastic scattering of the photon. This scattering is why the sky is blue, as the blue wavelengths are preferentially scattered. Another example of scattering is Mie scattering which was first described theoretically by Mie in 1908 and occurs

when the scattering object's size is comparable to the wavelength of the light. This is therefore a significant complication for IR spectroscopy as the commonly used wavelengths of 5-10 µm are similar to the size of many internal cell structures such as nuclei and other organelles.

Mie scattering is a disperse effect which is seen as a broad oscillation within the spectra, which both alters the position and intensity of the absorption peaks present at higher wavenumbers [111]. As the aim of using the FTIR instrument is to develop an accurate method for discriminating various tissue types and relating their spectral differences to features produced by known molecular species, it is therefore critical that the observed absorption peaks occur at the correct wavenumbers.

An algorithm developed by Paul Bassan *et al* [112, 113], was used to correct the spectra by approximating and then removing the distortion produced by the Mie scattering. The correction is an iterative algorithm based on extended multiplicative signal corrections (EMSC) and uses a Mie scattering approximation similar to Equation 4.2 [114, 115].

$$Q = 2 - (4/\rho)sin\rho + (4/\rho^2)(1 - cos\rho) \tag{4.1}$$

where Q is the scattering cross-section and

$$\rho = 4\pi a(n-1)\lambda \tag{4.2}$$

where a is the radius of the object it is scattering off, n is the ratio of refractive indices between the inside and the outside of the object and is the wavelength of the light. The correction is applied in an iterative process and so the spectra is improved in multiple stages. Once the final correction has been made for the Mie scattering affect it has been essentially removed from the spectra, producing a true representation of the absorption features of the sample [116].

Figure 4.1 shows an example spectrum taken from a raw FTIR image of OE 19 cells and the same spectrum once it has been processed by the correction algorithm.

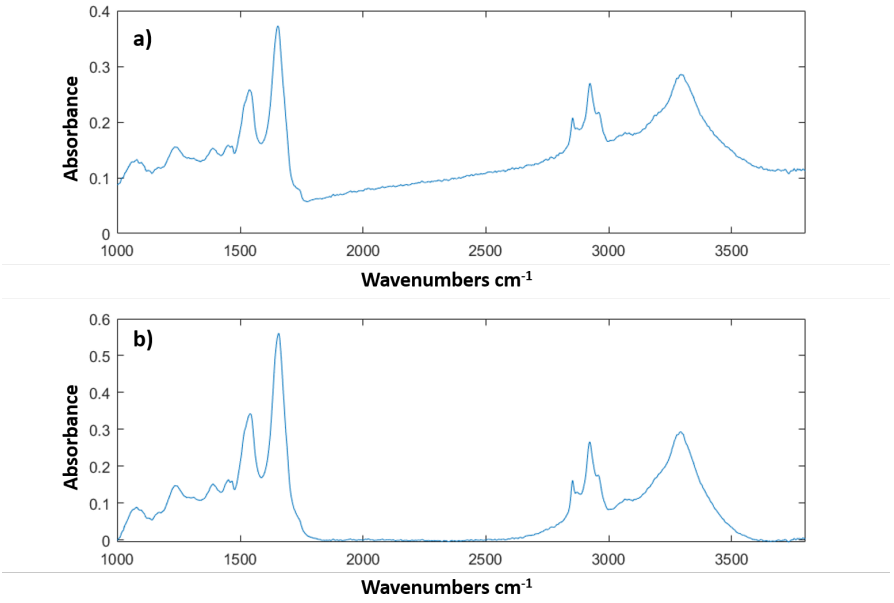

FIGURE 4.1: Figure showing the change in a spectrum after being corrected by removing the contribution due to Mie scattering, a) shows a raw spectrum taken from an FITR image of OE19 cells and b) the same spectrum shown in a) after being corrected.

4.2.3 Tissue labelling

The MA algorithm aims to identify characteristic features which differentiate the FTIR spectra of various sample types. It needs to be supplied with example spectra which have been correctly verified as belonging to a particular sample.

For FTIR images of tissue samples a code was written which allows a skilled user to label areas within an image as belonging to a specific sample type. To label the images correctly they rely on a detailed knowledge of the typical tissue structures and the use of adjacent slices which have been stained using H&E. The Barrett's oesophagus tissue biopsy images were labelled by Dr Olivier Giger (Institute of Translational Medicine, University of Liverpool). As it is often not possible to have an image which contains all of the desired sample types, spectra are collected from multiple images and combined so that a repository of many sample types

is formed. By combining spectra of the same sample type from multiple images it insures that the learning algorithm is as robust as possible, which ultimately produces a better classification model. Figure 4.2 shows examples of a labelled Barrett's oesophagus and oesophageal cancer sample.

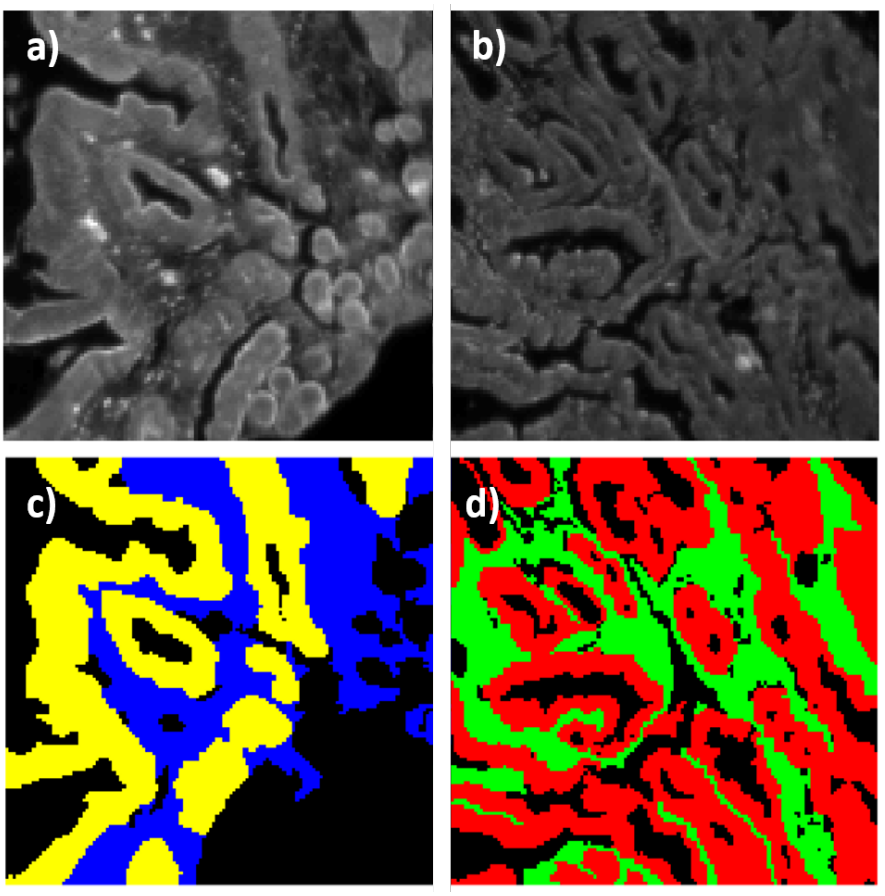

FIGURE 4.2: Figure showing the labelling of tissue biopsy images a) shows a representation of a FTIR image of Barrett's oesophagus tissue, b) is a FTIR image of a oesophageal cancer, c) shows the labelled areas of a) where yellow is the Barrett's epithelium and blue is the associated stroma and black is unknown tissue or blank slide, d) is the labeled ares of b) where red is cancerous epithelium and green canerous stroma.

The cell line images are simpler to label as the image ideally only contains spectra from a single sample type and therefore no manual labelling is needed. To remove spectra which are taken in areas of blank slide the total absorption across all the wavenumbers is calculated. Obviously areas of blank slide will have very little absorption and therefore spectra with a low total absorbance can be removed from the sample data sets. Figure 4.3 shows an example of a OE 19 cell line image where the spectra of the cell are clearly highlighted.

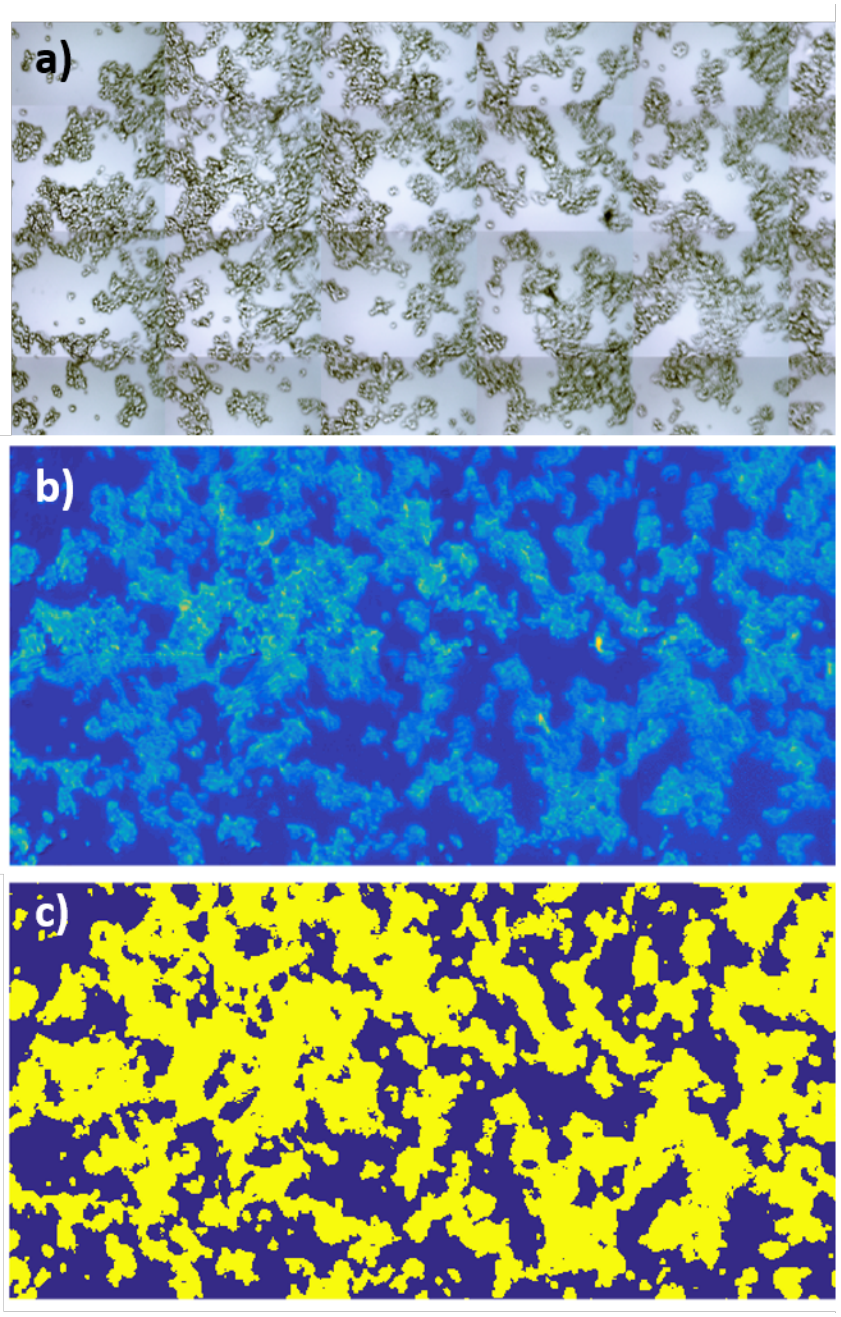

FIGURE 4.3: Figure showing the labelling of a cell line image, a) is an image of the cells taken by an optical camera, b) is a FTIR image of the same sample area as a) and c) is the labelled image where yellow is areas of cell and dark blue is areas of blank slide.

4.3 Metric analysis results and discussion

This section of the chapter will give an overview of the MA algorithm's performance at sample prediction as well as any possible novel insights the MA was able

to make. Results of analysing both biopsy tissues and grown cell cultures will be shown and contrasted as each may have their own tendencies and performances.

4.3.1 Tissue biopsy sample results

For this study two FTIR images were used to gather the datasets needed for MA. The first image was of a Barrett's oesophagus tissue biopsy and the second being an image of oesophageal adenocarcinoma. These samples were chosen for study, as a technique which is capable of quickly and reliably diagnosing if a patient with Barrett's oesophagus has begun to develop oesophageal cancer would be of great use for early detection purposes. From these FTIR images four tissue types were labelled by a skilled histopathologist and are shown in Table 4.1 along with their respective labels. The term 'associated' is used for the stroma samples to denote that they are simply taken from areas close to the adenocarcinoma and the Barrett's epithelium. Therefore AS isn't necessarily cancerous but it is expected that it's structure and chemical composition may have been altered by the surrounding AD. For brevity though the associated may be omitted, for example AS may be referred to a 'adenocarcinoma stroma'. The images were labelled carefully to ensure that there was minimal mislabelling, but as there was a limited amount of images available the labelling couldn't be too stringent as the MA requires at least a few hundred spectra to be robust.

Tissue type	Label
Barrett's oesophagus epithelium	BE
Barrett's associated stroma	BS
Oesophageal adenocarcinoma	AD
Adenocarcinoma associated stroma	AS

TABLE 4.1: A table showing the labels assigned to each sample type within the tissue biopsy study.

Figure 4.4 shows the average spectra for each of the four tissue types within the dataset and demonstrates that the spectra from different sample types are

still very similar with only relatively small variations between them. The average spectra does not give an indication as to how the absorption at a given wavenumber varies within a single dataset. For any classification method to be appropriate for tissue labelling it is critical to find reliable differences which will give a dependable method for classification.

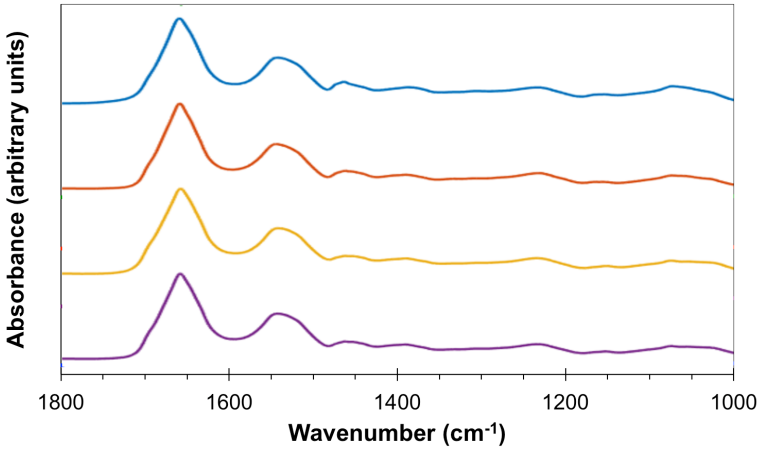

FIGURE 4.4: Figure showing the average of all the spectra for each of the tissue types within the dataset. This figure demonstrates that there is very little variation in the spectra between the various samples. Light blue is Barrett's epithelium, orange is Barrett's stroma, yellow is adenocarcinoma and purple is adenocarcinoma stroma.

The first and most obvious measure of the algorithm's performance to consider is it's success rate when using the optimal number of metrics for each tissue type. The success rate is the percentage of correctly labelled spectra within the testing dataset and is given individually for each tissue type. The success rates for each of the tissue types within the Barrett's and adenocarcinoma study achieved by MA is shown in Table 4.2.

Tissue type	Success rate %	# Metrics
Barrett's Epithelium	93	33
Barrett's Stroma	87	125
Adenocarcinoma	88	83
Adenocarcinoma Stroma	71	295

TABLE 4.2: A table showing the success rates of the MA algorithm at correctly labelling spectra of various tissue types and the number of metrics used in the optimal model.

The values shown in Table 4.2 are calculated using the optimal number of metrics for each of the tissue types within the analysis stage. The curves in Figure 4.5 show how the performance of the MA changes as a different number of metrics is used in the prediction model. Figure 4.5 helps demonstrate that the tissue types tend to behave differently with the maxima occurring at various positions, solidifying the need for individual optimisation. The curves show that there tends to be a significant increase in success over the first 10-15 metrics. In general the curves tend to peak and then plateau or begin to slowly decline as further metrics are used. The considerable improvement in the success rates is evidence that the voting system is beneficial for the best possible predictions. The values shown in Table 4.2 are found by taking the maximum value of the curves shown in Figure 4.5.

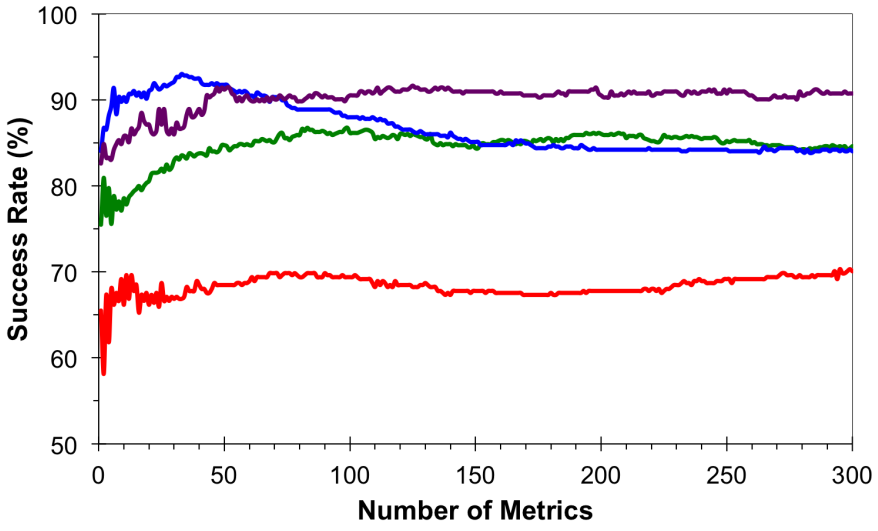

FIGURE 4.5: Figure showing the variation in the success rate for each tissue type as the number of metrics used in the prediction model is varied. Barrett's epithelium is blue, Barrett's stroma is purple, adenocarcinoma is green and adenocarcinoma stroma is red.

It is clear from Table 4.2 that as a sample classifier MA has performed well with an average success rate of 85% and it has worked especially well for the BE sample which had a success rate over 90%. The BS and AD samples both performed well with success rates of 87% and 88% respectively. The worst performing sample was

the AS which had a considerably lower success rate than the others at 71%, but this is still a reasonable value. These results are generated by the prediction of labels for spectra taken from real tissue images that were not used in the learning process. They are therefore a good indication as to how the classifying model would perform when labelling data taken from other FTIR images of similar samples.

For the 'blackbox' classifier needed for tissue diagnostics the success curves are all that is required to ascertain the performance of the model. The following figures are designed to study which wavenumbers have been highlighted as important for discrimination. The most useful plot for assessing all the wavenumbers simultaneously are the butterfly images which are shown in Figure 4.6. As the axes of a butterfly image represent each of the $\bar{\nu}$ values used in the ratio pair that define each metric, the score for a given metric can be determined by finding the point at which the two wavenumbers intersect. The scores for all metrics analysed for a given sample are represented in a single image, which allows for a user to quickly get a grasp of which wavenumbers are important. The butterfly plots also allow the user to visually compare how certain wavenumbers preformed for each of the samples. The plots have an associated colourmap which represents the performance of a metric with dark red being a very good metric and dark blue being very bad.

The butterfly plots in Figure 4.6 show that ratio pairs perform differently from sample to sample, with the highest scores occurring in different areas of the butterfly plot. This demonstrates that each tissue type tends to have unique wavenumbers which are able to distinguish them from the other samples. The areas which show good performance occur over several pixels which indicates that the MA is sensitive to small features within the spectra rather than specific wavenumbers. This is expected as the spectra don't demonstrate any sharp features with the spectral resolution of 6 cm^{-1}. Each of the butterfly plots show multiple areas which achieve a good score, which implies that the spectra have multiple features within the 1000-1800 cm^{-1} range that can be used for classification. This infers that there are multiple chemical variations between the tissue samples, which would obviously be of interest to researchers.

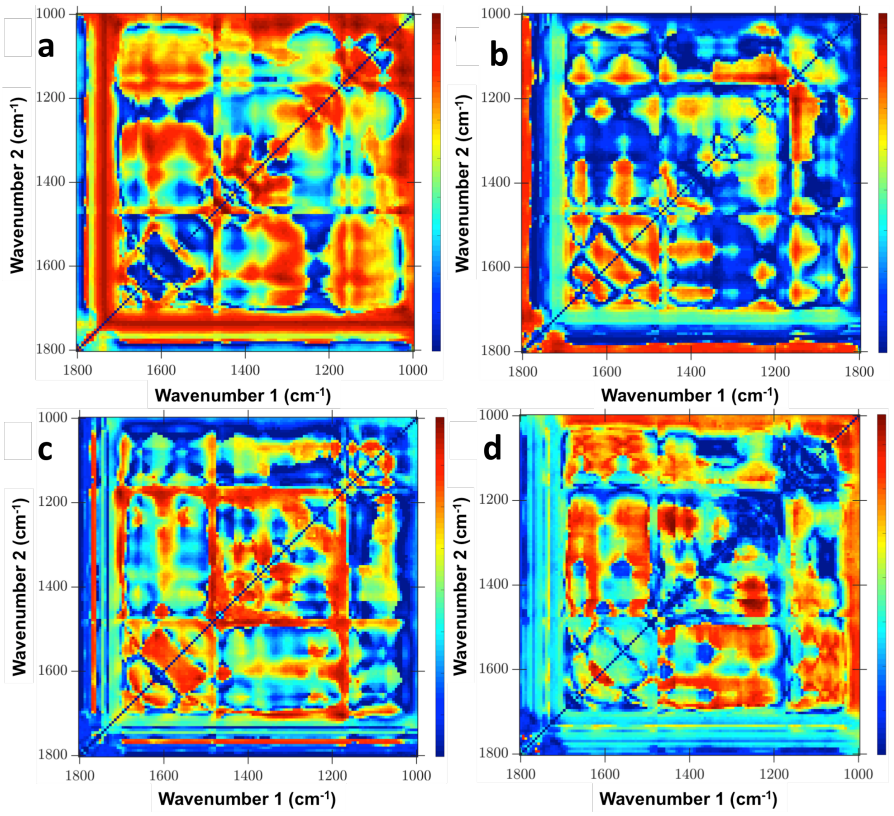

FIGURE 4.6: Figure showing the butterfly plots of the Barrett's and adenocarcinoma samples, a) Barrett's Epithelium, b) Barrett's Stroma, c) Adenocarcinoma d) Adenocarcinoma Stroma.

Although the butterfly images are useful in describing the general importance of all the wavenumbers it doesn't obviously highlight which wavenumbers are the best and also which are used in the optimal metric set in the classification model. To do this a discrimination plot is used, which is a plot that tallies how many times a wavenumber is used within the ratio pairs of the top metrics for each tissue. The more times a wavenumber is used in the best metrics the more important that wavenumber is deemed to be. The discrimination plot for the top 5 metrics is shown for each tissue sample in Figure 4.7.

By studying the wavenumbers used in the top 5 metrics for each sample, it is clear that in general each sample uses a unique set of wavenumbers to classify the tissue samples. This is true apart from $\approx$ 1460 cm^{-1} which is used by both the metrics for Barrett's epithelium and adenocarcinoma. Multiple samples using the same wavenumber is not detrimental as it simply indicates that the wavenumbers are key for sample discrimination and may ultimately help produce a simple device

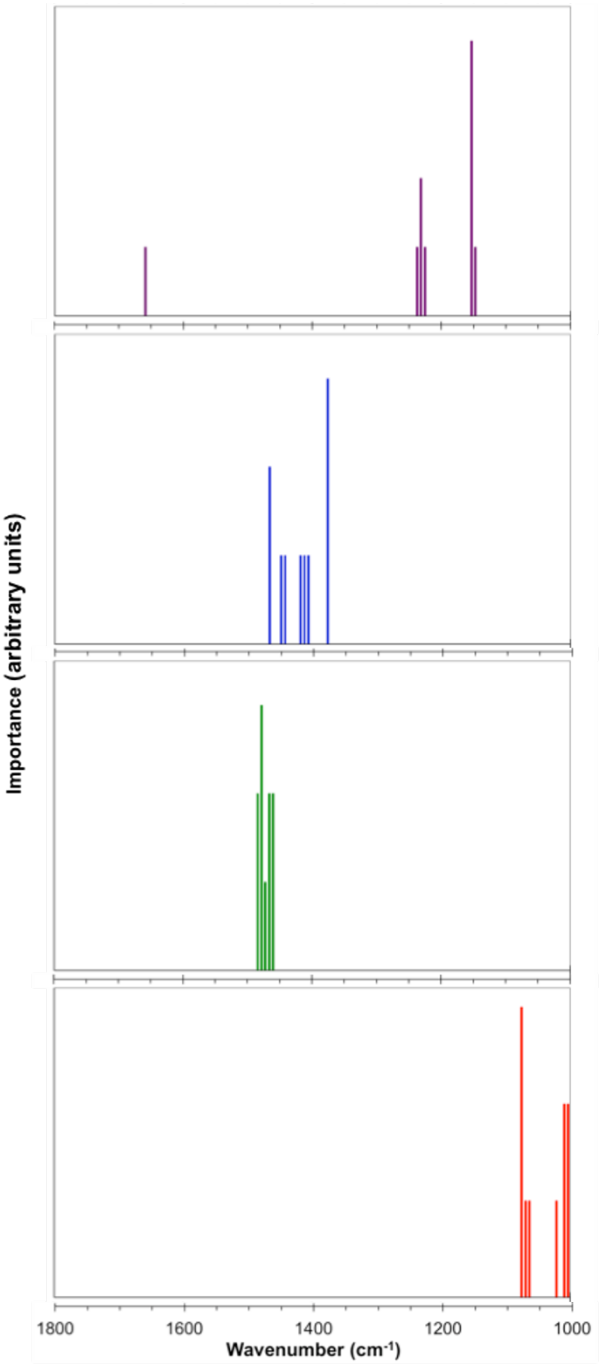

FIGURE 4.7: Figure showing the discrimination plots, which indicate the importance of every wavenumber used in the top 5 metrics for each tissue type. Purple is the Barrett's stroma, blue is the Barrett's epithelium, green is the adenocarcinoma and red is the adenocarcinoma stroma.

capable of diagnosing multiple tissue samples with only a few wavenumbers being needed.

4.3.2 Tissue biopsy sample discussion

By assessing the success rates of the MA algorithm it is clear that it is capable of discriminating well between the tissue samples. Insight into which wavenumbers were important for discrimination can also be assessed from the butterfly and discrimination plots.

The MA performed significantly worse for AS compared to the other samples, so to better understand the MA's performance the labels given by the MA can be studied in further detail. Table 4.3 shows the percentage of how each tissue type was labelled within the dataset using the optimal number of metrics. Most of the tissue types show a large portion of correctly predicted spectra with a considerably smaller proportion of incorrect predictions. This is true apart from for the lowest performing tissue type which was AS which shows 20% of the AS testing dataset was labelled incorrectly as AD. This is almost a third the size of the number of correctly labelled spectra, which is not demonstrated to the same extent in any of the other tissue types. The reasons for such a considerable mislabelling will be explored further.

Tissue type	BE%	BS%	AD%	AS%
Barrett's Epithelium	93	6	1	0
Barrett's Stroma	6	87	2	5
Adenocarcinoma	2	3	88	7
Adenocarcinoma Stroma	5	4	20	71

TABLE 4.3: A table showing the proportion of each datasets were labelled.

There are many possibilities as to why AS performed worse than the other samples types. The first is simply that the MA is unable to properly characterise the AS sample. This is unlikely though as it achieved a considerable success rate of over 70%, which would be unlikely if the prediction algorithm had no grasp on the key characteristics of the spectra. Also if the MA was struggling to distinguish between AD and AS it would be expected that the poor performance would be

seen in both of the sample's success rates. This is not the case though as the portion of AD incorrectly labelled AS is 7% which is only just larger than that of BE and BS.

A more likely possibility is that the AS tissue varies significantly in nature at various spatial positions within the sample. This is possible as stroma is a general term given to the connective tissue which borders well-defined tissues which have a dedicated role such as epithelial or muscle tissue. If the AS chemical content varied significantly enough so that there was essentially multiple tissue types present in the same dataset, multiple distributions may be present with the training stage for a single ratio pair. This would result in a poor PDF fit as the distributions are always fitted with only a single Gaussian, the PDF would therefore not represent the data as well as it normally does. It is not inconceivable that the AS tissue may vary as it only associated to the AD which is present near it. This means that the AS is from an area which has AD present but is not itself defined as being either healthy or cancerous. AD is liable to changing the structure of the surrounding stroma [117], so it may be that within the AS dataset there is a mixture of healthy and cancerous stroma or cancerous stroma which have been altered to varying degrees. Some of the AS may have been altered so significantly by the neighbouring AD that some of it's spectral features appear more like the AD than that of normal AS, which would explain why there is a significant portion AS mislabelled as AD.

It is also possible that some of the AD tissue was simply mislabelled AS by the histopathologist, which is possible as it is much harder to distinguish AS from AD than it is to tell the difference between the much more structured BE and BS. It is interesting to note that mislabelled spectra will have a varying impact on the different stages of MA. The training stage is actually quite robust as long as the portion of mislabelled spectra is not so overwhelming that multiple significantly sized distributions are produced. The success rates calculated in the testing and analysis stages are much more sensitive to the mislabelled spectra as they assumes that all the predefined labels are correct. For example, if two samples are studied (A and B), but 10% of A's spectra are actually taken of sample B, both the training

and testing datasets for A will contain $\approx 10\%$ of B spectra. The characterisation of A will still be successful as the mislabelled spectra will not greatly affect the fit of the PDF to the majority of correctly labelled A spectra. The result of the training stage is therefore PDFs which can describe the spectra well and discriminate both of the samples. The good performance of the training stage doesn't matter when calculating the success rates in the following stages as the mislabelled spectra within the training set of A will be predicted correctly as being B, as the MA was still able to characterise the data well. But as the predefined labels are assumed to be true the correct prediction will be counted as being incorrect as the MA believes all the spectra within the dataset for A must be A. This would therefore lower the apparent success rate even though the actual label predictions had been correct. The problem of mislabelling could be limited by being very selective when labelling the FTIR images, which would be ideal but couldn't be done in this study due to the limited number of images.

The reasons just explained may also be responsible for the other tissue types incorrectly labelling their spectra. Clearly though this effect is much smaller or the number of mislabelled spectra is considerably lower for the other tissue samples when compared to AS. From reviewing the results shown in Table 4.3 the MA demonstrates it's potential as a sample classifier and by knowing the issues associated with the AS datasets, the poor performance of the MA to classify AS can be explained and possibly rectified in future experiments. The ideal solution would be to rerun the experiment with much larger datasets from multiple images, which are meticulously labelled to ensure minimal mislabelling.

Although it is hard to make concrete statements on the chemical differences between the tissues samples from such a small study, the wavenumbers found to be important by the MA can be compared to previous IR studies to ascertain if they coincide. A paper which utilised FTIR to image oesophageal tissue [72] showed that 1450-1465 cm^{-1} are significant wavenumbers which are attributed to both proteins and lipids. These wavenumbers are found to be highly significant for both BE and AD as demonstrated in the discrimination plot, Figure 4.7, indicating that the relative levels of both the lipids and proteins may be a discriminator of

the tissues. In the same paper 1237 cm^{-1} was noted as a strong discriminator between Barrett's and malignant tissue and is attributed to nucleic acids. The region of the IR spectrum around 1237 cm^{-1} was also found to be an important wavenumber by MA as it occurs multiple times in the optimal metric sets for both AS and BS.

Another paper using IR spectroscopy to study cervical cells [118] found significant variations between healthy and cancerous tissue at 1155 cm^{-1} which is associated to the C-O stretch within proteins. Although this was originally discussed for a different cancer 1155 cm^{-1} was found to be an important discriminator for the BS. This may therefore be indicating something which may relate to cancer in general, which would obviously be interesting. Important wavenumbers are observed in both the stroma sets by MA in the 1545 cm^{-1} region. This region is associated with the Amide II band found in proteins, which has also been flagged as a previous potential discriminator.

Finally changes in the absorption bands at 1080 cm^{-1} were observed in [72] and attributed to the variation in nucleic acids between cancerous and healthy tissue. This wavenumber is found to be important by the MA as it it is appears often within the AS optimal set. Table 4.4 summarises the important wavenumbers and the molecular species which have been previously associated to potential discriminators.

Wavenumber cm^{-1}	Main chemical contributor	Sample	Reference
1080	Nucleic acids	AS	[70, 72, 118]
1155	Proteins	BS	[72, 118]
1237	Nucleic acids	AS, BS	[70, 72]
1450-1465	Proteins and lipids	AD, BE	[72]
1545	Amide II	AS, BS	[72]

TABLE 4.4: A table showing wavenumbers that the metrics analysis deemed important for discrimination for each tissue sample (BE is Barrett's epithelium, BS is Barrett's stroma, AD is Adenocarcinoma and AS is Adenocarcinoma stroma). The main molecular species associated with IR absorption at each particular wavenumber is stated along with the reference to the supporting literature.

The study of the two tissue FTIR images has demonstrated the capabilities of MA not only as a predictive tool but also it's ability to give insight into what are the key distinguishing chemical differences between the tissue samples. Although this study only contains a limited amount of data the MA, with no built in bias to known important wavenumbers for biological classification, was able to find multiple wavenumbers which have previously been determined to be significant for tissue discrimination. As the majority of the important wavenumbers shown in the discrimination plots can be assigned to known molecular species common in biological samples it gives credibility that MA is extracting relevant and important chemical information within the IR spectrum and is not finding differences based on other factors, such as artefacts within the FTIR images.

The tissue study has highlighted some of the dangers with MA, as it is sensitive to the quality of the data used in the learning process to both characterise the sample but to also give reliable assurances on it's performance. This can be hard to guarantee in tissue images as they are often very complex with subtle borders between one type of tissue and the next, but could be remedied by carrying out a larger experimental study where the spectra are taken very carefully.

4.3.3 Cell line sample results

As the study of tissue biopsies showed that MA had promise as a classification tool, further experiments were needed to continue testing it's potential. As the quality of the labelling within the datasets used in the learning process was found to be critical, the experiment was designed to fundamentally overcome this issue. Each cell line sample contains cells of only a single type, which were grown on a CaF_2 slide to allow them to be used in IR instruments. If the slide was kept free from contamination none of the spectra collected for the MA could be mislabelled. A total of five cell lines were imaged in this study with various oesophageal carcinoma samples (OE 19, OE 21 and OE 33) and two myofibroblast samples (173/1 and 173/5). The cell line samples are shown in Table 4.5 with their associated labels.

Cell line sample type	Label
Oesophageal adenocarcinoma (oesophageal gastric junction)	OE 19
Oesophageal squamous cell carcinoma	OE 21
Oesophageal adenocarcinoma (lower oesophagus)	OE 33
Cancer associated myofibroblast	173/1
Non cancer associated myofibroblast	173/5

TABLE 4.5: A table showing the cell line samples and their associated labels.

It was found when reviewing the FTIR data that there were considerable artefacts present within the OE 33 sample image, which meant it could not be used with the MA algorithm as it had negative effects on the performance for all the samples. This is because the training stage essentially compares all the PDFs within a metric. If a dataset is behaving sporadically the PDF wouldn't represent the true chemical content and so would behave differently, this would have a detrimental effect on the assessment of the other samples as all the PDF are compared to the OE 33. As the OE 33 dataset included artefacts, wavenumbers found to be good discriminators would not necessarily relate to chemical features and may instead be highlighting the presence of the anomalous data. Therefore the assessment of the important wavenumbers for discrimination would yield very little. The single OE 33 image contained multiple FTIR scans as it was a mosaic but the artefacts were found throughout and therefore the OE 33 data had to be fully removed from the cell line study.

The average spectra of the four cell lines used in this study are shown in Figure 4.8, as with the tissue study it is clear that there is little variation between the average spectra.

The 173/1 average spectrum is visibly different at around 1025 cm^{-1}, but assumptions of this being a good discriminator cannot be made at this point as the average spectra do not account for the magnitude of the variation within the 173/1 spectra. As with the tissue studies, the success curves and the success rates using the optimal number of metrics is the first stage for assessing the performance of the MA. The success curves for the cell lines are shown in Figure 4.9.

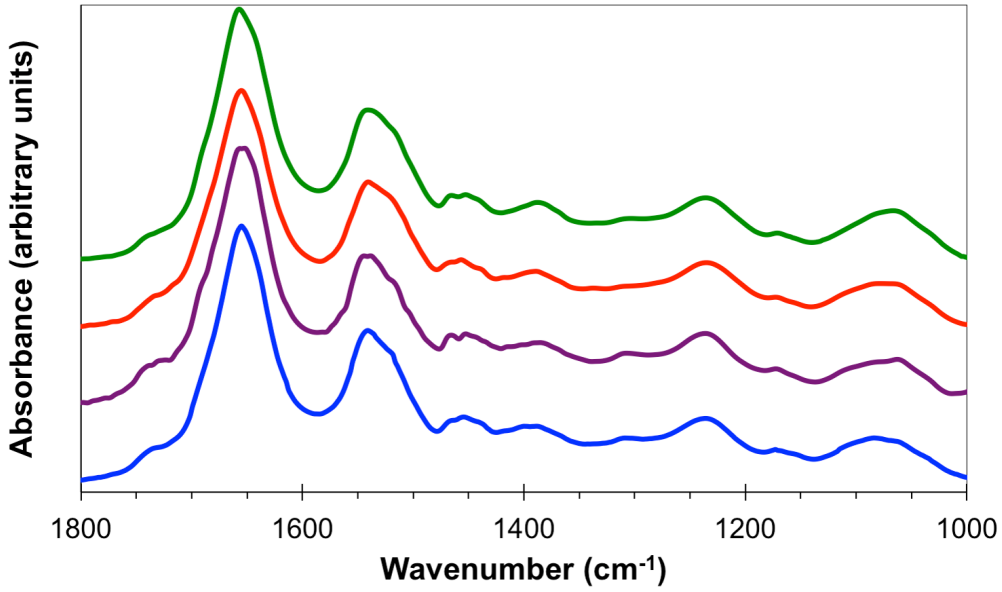

FIGURE 4.8: Figure showing the average spectra of each of the cell lines used in the learning process. Green is OE 19, red is OE 21, purple is 173/1 and blue is 173/5.

Figure 4.9 shows that all the samples perform well with OE 19, 173/1 and 173/5 all achieving 90%+ success rates. The most dramatic increase in the success rate seen in both the tissue and cell line study is demonstrated by the 173/5 sample which at one point has a very low success rate of $\approx 45\%$ and increases to around 90% when the number of metrics is increased. As the metrics are used in order of their ranking with the best being first, this dramatic increase implies that none of the metrics were individually very capable of high prediction rates for 173/5 and the classifier model relied heavily on the collaborative affect of the voting system.

The success rates of each cell line when using the optimal number of metrics are shown in Table 4.6.

Tissue type	Success rate	# Metrics
OE 19	97	2
OE 21	81	1
173/1	92	64
173/5	91	24

TABLE 4.6: A table showing the success rates of the MA algorithm at correctly labelling spectra of various cell line samples and the number of metrics used in the optimal model.

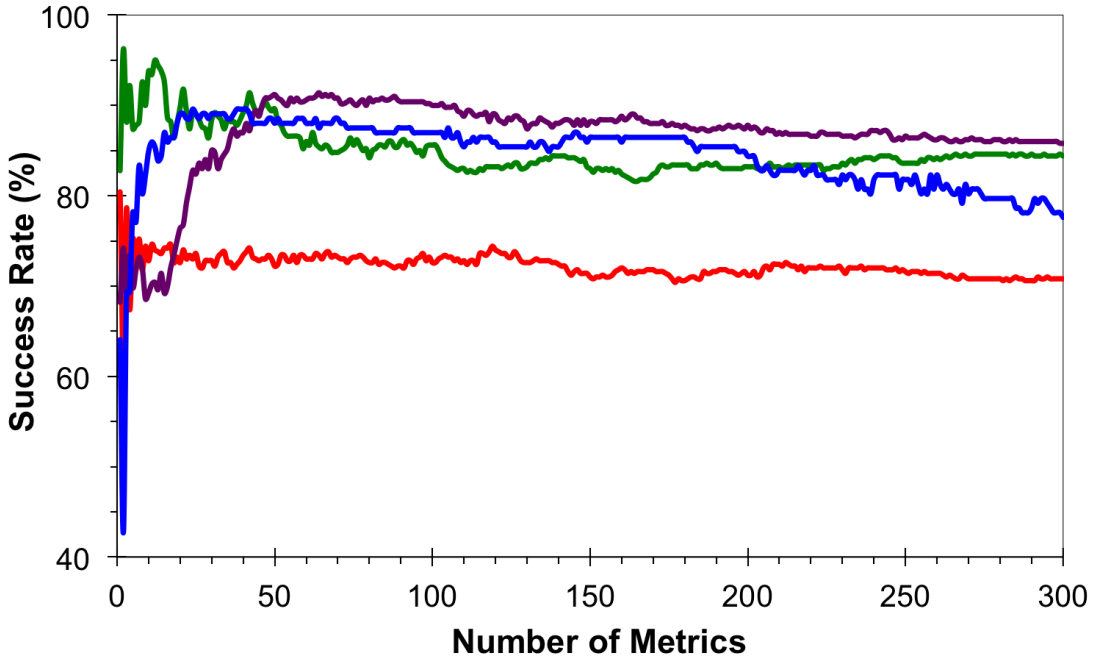

FIGURE 4.9: Figure showing the variation in the success rate for each tissue type as the number of metrics used in the prediction model is varied. Green is OE 19, red is OE 21, purple is 173/1 and blue is 173/5.

The average success rate of the cell line study was greater than that of the tissue study at 90%. OE 19, 173/1 and 173/5 performed particularly well with success rates higher than 90% and OE 19 achieved a success rate of 97% which is very high given that the spectra used to gain this value were not used in the training stage. The worst performing sample was OE 21 with a success rate of 81%, which is still a very good rate of labelling. The high success rates again signify that the MA is performing well and is able to discriminate the cell line samples well. The butterfly images generated by the MA for the cell line samples are shown in Figure 4.10.

As with the tissue study the butterfly images show multiple features for each of the cell lines indicating that several areas of the IR spectra can be used to distinguish them. The discrimination plot for the cell lines is shown in Figure 4.11. The discrimination plot again demonstrates that the wavenumbers in the very top metrics tend to be unique between the samples. There are interesting

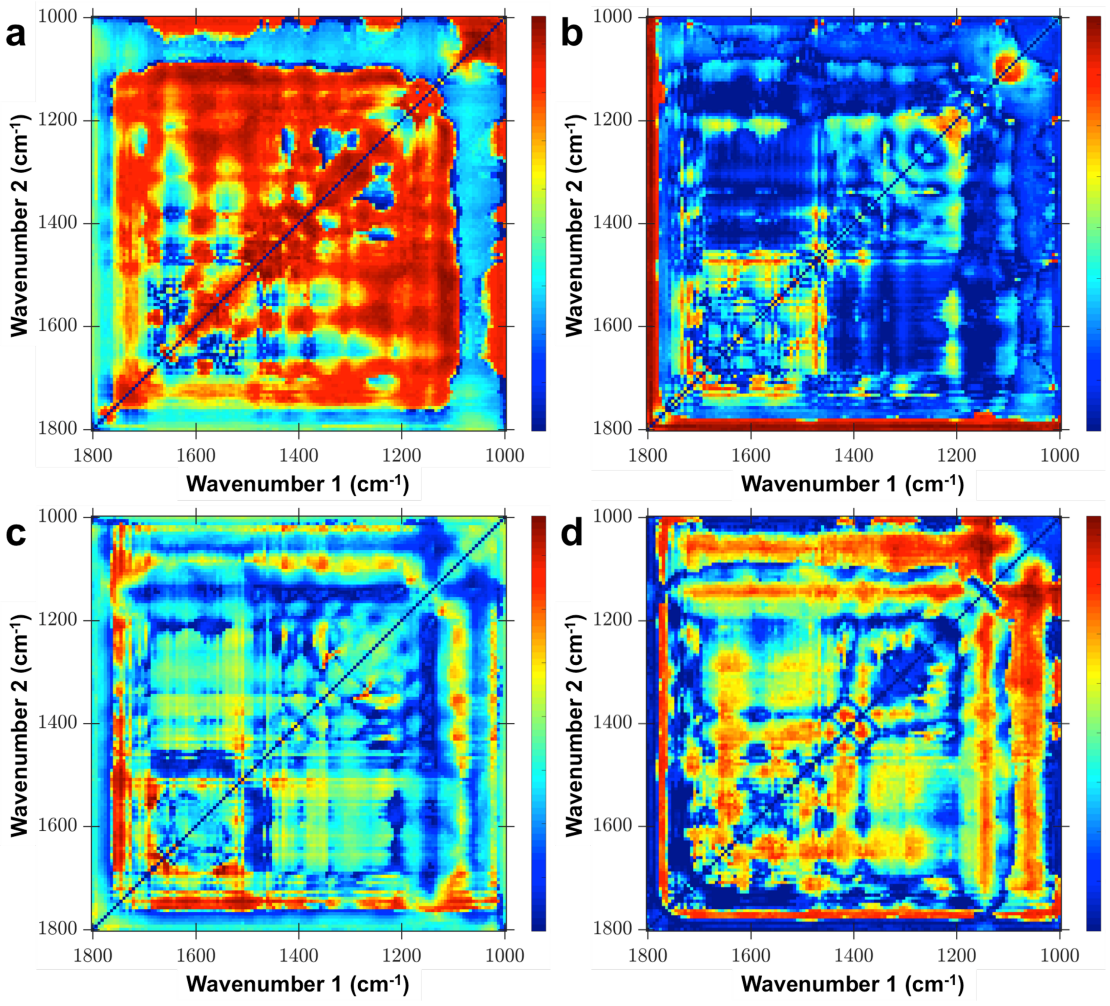

FIGURE 4.10: Figure showing the butterfly plots of the cell line samples, a) OE 19, b) OE 21, c) 173/1 d) 173/5.

features such as an area of importance for 173/1 (purple) at 1510 cm^{-1} which occurs in between the two significant areas of interest for OE 21. There is some overlap though as both OE 21 and 173/5 use 1475 cm^{-1}.

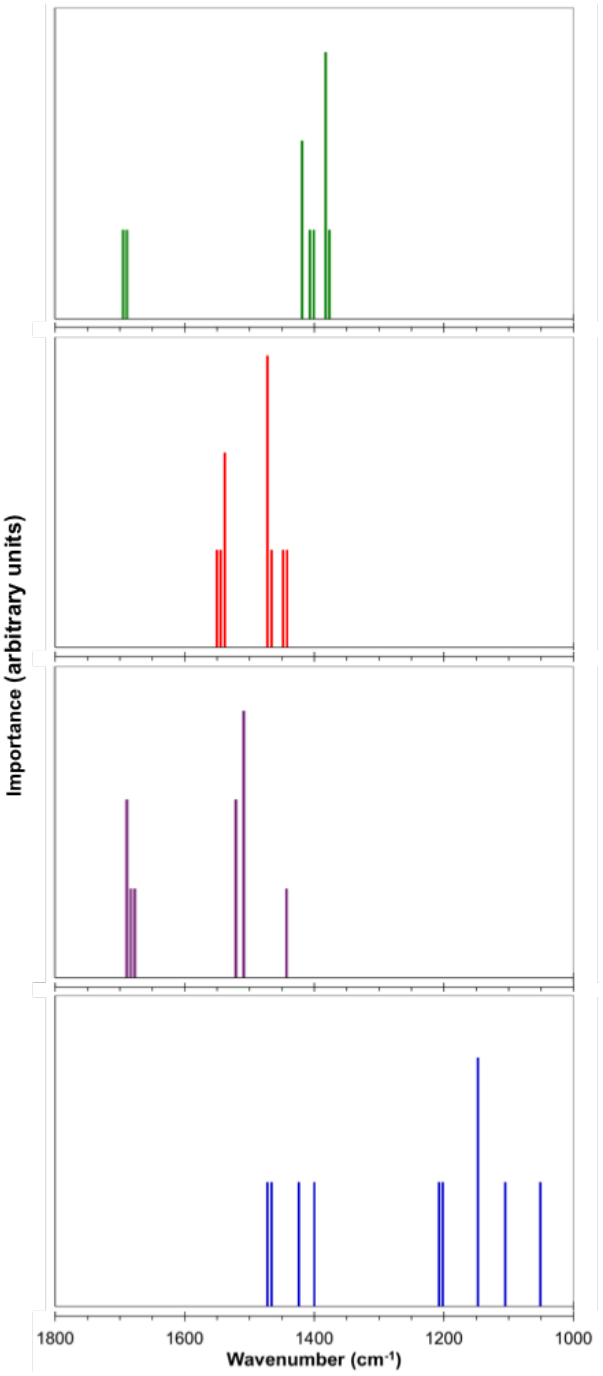

FIGURE 4.11: Figure showing the discrimination plots, which indicate the importance of every wavenumber used in the top 5 metrics for each cell line sample. Green is OE 19, red is OE 21, purple is 173/1 and blue is 173/5.

4.3.4 Cell line sample discussion

With the high success rates shown for the cell line study the MA has again demonstrated that it is a powerful tool capable of discriminating between the various samples. What is significant within the cell line study is that currently this is

the first example of an automated system being able to distinguish the two myofibroblast samples taken from the same patient [119]. The performance of both the myofibroblast samples were very high, which given the current difficulty of discriminating the two indicates the potential for MA.

The samples used in the cell line study are very similar to each other like in the tissue study, which is demonstrated in Figure 4.8 by the little variation between the average spectra. The performance of the MA has improved from the previous study which indicates that the lack of mislabelling appears to have helped build a better and more accurate classifier model.

Another interesting difference between the cell line and tissue biopsy studies is that there are considerably less metrics used in the cell line classifier model. The tissue model used 135 metrics on average while the cell line study only used 23, with OE 19 and OE 21 only needing 1 and 2 metrics respectively. This lower number of metrics may indicate that the spectra within the cell line datasets were more homogeneous, resulting in better classification. This is because a smaller number of metrics are needed to produce a sufficiently robust model which is capable of handling the effects of the spectral variance.

A further comparison which can be made between the two studies is comparing the metrics used by OE 19 and AD. OE 19 is a cell line of oesophageal adenocarcinoma cells which originate from the oesophageal gastric junction. AD is spectra taken from areas of tissue which had been labelled as oesophageal adenocarcinoma by a trained histopathologist. The OE 19 and AD samples are the two samples which in principle are the most akin between the tissue and cell line studies. It would therefore be interesting to compare the performance and results of the MA on these two samples.

Both achieved a high level of discrimination with success rates of 97% and 88% respectively for OE 19 and AD. As OE 19 only used 2 metrics and AD used 83 it wouldn't be fair to simply produce a discrimination plot which shows the important wavenumbers in the the optimal model. Instead Manhattan plots can be used to visually show which wavenumbers are used in the top 100 metrics for

both the OE 19 and AD samples, the Manhattan plots for both are shown in Figure 4.12.

In a Manhatton plot the y-axis displays the metric number and the x-axis describes the wavenumber. At a value of n on the y-axis, each of the wavenumbers which were used in the top n metrics are displayed. Therefore at $n = 1$ only the wavenumbers in the very best metric are displayed, likewise if $n = 50$ all the wavenumbers used in the top 50 metrics are plotted. This is why the structures tend grow as more metrics are studied as each line contains all the points which were in the previous. The colours denote whether the wavenumber used was either $\bar{\nu}_1$ (red) or $\bar{\nu}_2$ (blue) in the ratio pair of the metric.

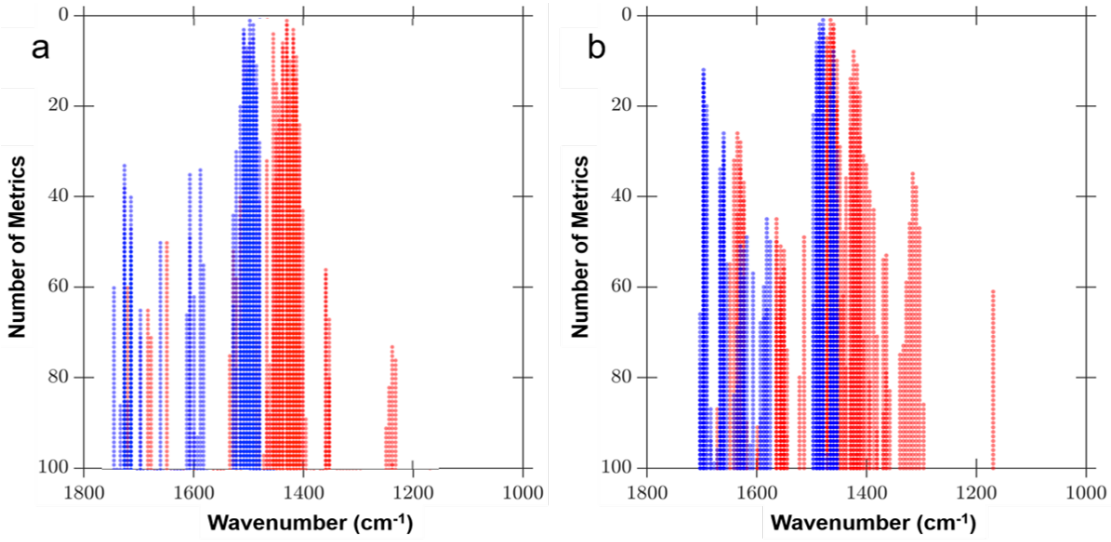

FIGURE 4.12: Figure showing the Manhatton plots for a) OE 19 and b) AD.

The Manhattan plots are useful to visualise which wavenumbers are important for a given sample. They also allow for the easy comparison between two samples. As the OE 19 and AD samples should similar in chemical composition as they are both oesophageal adenocarcinoma, it is possible that the important wavenumbers determined by MA may also be similar.

Clearly in Figure 4.12, the two Manhatton plots are very similar generally as they both have two main features centred at ≈ 1420 cm^{-1} and 1465 cm^{-1}. They both show little importance in the wavenumbers which are smaller than 1300 cm^{-1} and some features in the 1500-1700 cm^{-1} area, although the AD sample

seems to find these more important than OE 19. They even both show the same general distribution as to which of the wavenumbers are used as $\bar{\nu}_1$ (red) and $\bar{\nu}_2$ (blue). As $\bar{\nu}_1$ tend to be used for the lower wavenumbers and $\bar{\nu}_2$ tends to be the higher wavenumbers. There are some differences within the Manhatton plots but this is to be expected as the these wavenumbers were found to be important discriminators against the other samples within their individual study. As these other sample types were different it is clear as to why the wavenumbers within the metrics may vary. For example, the tissue study includes Barrett's samples which aren't present in the cell line study and therefore a metric which is helpful at distinguishing Barrett's epithelium (BO) and adenocarcinoma (OE 19 and AD) would be useless in the cell line model as it has not been trained to separate OE 19 and BO.

The significance of this comparison is that the MA has found common core discriminators between two samples which although they are both adenocarcinoma they are also different in many ways. OE 19 is a cell line sample which has been grown on a slide, whereas the AD sample has been taken directly from a patient as a biopsy. These are samples taken from different people and the results shown here were generated in two completely separate studies which involved different sample types. The spectra used in the MA learning process were collected in different images and were taken years apart. Many of the key wavenumbers can be attributed to known absorption bands of biological molecules, which would be expected as certain molecules are known to be influenced by cancer. This is significant as MA has no inherent bias towards wavenumbers which are known for discrimination. This comparison indicates that cell line samples which are easy to control and label could potentially be used to label the same sample types in tissue biopsies. This would remove the issues of extracting the learning datasets from tissue images, while still being able to accurately diagnose tissue samples.

What is also interesting is that although some of the important wavenumbers coincide with previously documented discriminators which is to be expected, there are wavenumbers the MA has highlighted have not been previously flagged. This is in some ways more exciting as they may lead to new insights and interpretations

which have not previously been made. This may lead to an increase in the understanding of the mechanics occurring within cancer, which may ultimately help diagnostics and treatments.

4.3.5 Metrics analysis and random forest comparison

As MA is a newly developed form of ML it would be prudent to assess it's performance against that of another commonly used classifier. It has already been compared to CA by Timothy Craig and was found to be more appropriate for spectra discrimination. One such ML technique is called random forest (RF) which has been used frequently to classify a wide variety of biological samples [120–122]. To compare the two techniques a MATLAB version of an RF classifier program was acquired from an online source [123]. RF is a supervised algorithm which works on a similar principle to the MA, in that it aims to find features within example sets to build a classification model. The original RF program was slightly altered to work in a similar way to the MA so that it randomly separates the same datasets into a learning set and a testing set. This ensures that like the MA the success rates are a fair approximation as to the programs performance as the test spectra aren't used within the learning process.

To compare MA and RF they were assessed by both their success rates and the processing time. They were tested using the cell line samples and the tissue samples. The main time dependence for the MA is the number of wavenumbers studied within the training spectra. As a metric is formed for every possible wavenumber pair the the number of metrics which need to be generated and assessed (which is the majority of the processing time) is proportional to the square of the number of wavelengths. Because all the metrics are always assessed the time taken is very predictable. When using RF the user sets the number of trees to be used which are classifiers analogous to metrics. The more trees used, generally the better the classification model should perform as there are more classifiers available, but the processing time will also be longer. RF is not like MA though in that fact it is an example of an optimiser method. It runs through multiple iterations adjusting

parameters in it's model until no improvement appears to be gained at which point the code stops. This means the processing time can often vary greatly depending on the data it is processing. This is demonstrated by the variation in processing times for the cell line and tissue studies. The results of both the MA and RF for the cell line is are shown in Table 4.7.

	Metric analysis	Random forest	Random forest
Time taken (s)	96	1254	31
Number of trees	N/A	500	10
Resolution cm^{-1}	6	20	20
OE 19	97	95	92
OE 21	81	54	51
173/1	92	96	92
173/5	90	10	16

TABLE 4.7: A table showing the relevant settings and results for both metric analysis and random forest for the cell line data.

To insure that the results are fair both the MA and RF were used with the same version of MATLAB and ran on the same computer under similar conditions. The RF was originally ran using the same resolution as the MA (6 cm^{-1}), but using this resolution with a reasonable number of trees resulted in the RF program taking an impracticable amount of time to process. The resolution of the spectra was increased to 20 cm^{-1} which made the computational times much more akin to MA. To demonstrate the variation in the performance of the RF using a different number of trees, Table 4.7 shows the success rates and processing time for 10 and 500 trees.

Table 4.7 shows that when using only 10 trees the RF and MA perform similarly for both the OE 19 and 173/1 datasets, with both methods performing well with success rates of over 90%. The clear distinction between the two techniques is found with the OE 21 and 173/5 samples. Even when the RF program uses 500 trees, which results in a processing time over 13 times longer than that of the MA it is still out performed for the OE 21 and 173/5 samples.

The results of comparing the MA to RF when using the tissue biopsy sample are shown in Table 4.8. It is clear that the code is able to process the tissue data quicker than the cell line as a scan using 5000 trees only took 648 seconds. A scan using 500 trees for the cell line data took over twice this amount of time.

	Metric analysis	Random forest
Time taken (s)	246	1254
Number of trees	N/A	648
Resolution cm^{-1}	4	20
BO	93	82
BS	87	88
AD	88	88
AS	71	54

TABLE 4.8: A table showing the relevant settings and results for both metric analysis and random forest for the tissue biopsy data.

Table 4.8 shows that MA tends to generally outperform the RF classifier in both success rates and also processing time although there is less of a difference between them compared to the cell line study.

4.4 FTIR conclusion

The aim of studying the FTIR was not to asses its capabilities as an instrument, which has already been demonstrated in a very large number of wide ranging experiments. The aim was instead to take images of oesophageal samples and ascertain if sufficient information could be gathered so that a ML technique could be used to classify the various samples. The classifier would then be able to use the discriminators it had found to predict the sample type of data which it has not encountered before.

A variety of techniques were considered but because they often weren't appropriate or didn't fulfil all the desired requirements MA was designed. The MA algorithm was developed to do this and was shown capable of studying both cells and tissues samples, finding reliable characteristic differences between them. The MA achieved an average success rate of 85% and 90% for the tissue and cell line respectively. This demonstrates that MA has been able to meet the need for a 'blackbox' diagnositc tool.

A significant result of the cell line study was that MA was able to successfully distinguish spectra from cancer associated myofibroblasts (173/1) and non cancer associated myofibroblasts (173/5), which are very hard to differentiate when using standard optical microscopes and dyes. This demonstrates that the MA program is be able to accomplish things which are not possible with current techniques.

The wavenumbers found to be important by the MA for discriminating the tissue samples have been compared to other IR spectroscopy experiments. The results showed that some of the wavenumbers found to be important by the MA agree with the previous studies. This is important as there is no inherent bias in the program to absorption bands which are related to biological molecules, the MA has therefore been able to extract key chemical information. This was the second aim for the MA, which was to provide detailed information as to the relative chemical compositions of biological samples and then to present the wavenumbers which are able to separate the samples. By studying these discriminators it may

be possible for a greater insight into the samples inner workings which in turn will help with diagnostics and treatments. What is also interesting is that there are wavenumbers found to be significant which haven't been flagged in previous studies. This may be the first indication of discriminators which have not yet been appreciated by the scientific community. The MA is therefore able to meet both the needs of a clinician and a researcher which was the ultimate aim for this study.

Finally MA was compared to the established ML technique of RF to see how it performs. The MA was found to outperform the RF in both it's ability to predict the correct labels for spectra but also the processing time required for the learning process.

Although the MA results are promising, it is important to note that it has only been tested on a relatively small sample set, which is far to small to greatly influence the medical community. The next step for MA is therefore to continue the studies with larger datasets which are of a scope that is more akin to medical trials. As MA is a general purpose classification program it is appropriate for studying other cancers and diseases.

Chapter 5

SNOM study

This chapter will focus on the results of the SNOM experiments with the IR-FEL attached to the ALICE accelerator. The main aim of this chapter is to assess the potential for SNOM to be used in biological studies to gain images with resolutions that were previously not possible. The SNOM experiments have been ongoing for several years, with which I have been a major contributor of, with the many images being collected and assessed by the Liverpool SCAnCan group to discover problems which can be corrected to improve the performance of the SNOM for the subsequent run. As the results and the implemented improvements, both of which I was involved in, have been discussed in detail by Timothy Craig in his PhD thesis [74], this chapter will focus on the key results gained in the final run which I led, as they will give the greatest insight into the current state of the SNOM instrumentation. For clarity all the data shown within this chapter was collected by myself.

5.1 SNOM background

Due to the constant pressure to improve the performance of imaging devices, the drive to develop high resolution instruments has been growing as many other traditional imaging devices have begun to reach their natural limits imposed by

far-field physics. This has been especially true for IR spectroscopy techniques as they have the possibility to provide information on the chemical contents of the biological samples but have often suffered with poor resolutions compared to the standard optical microscopes which use visible light. Due to the heavy requirements needed to operate a SNOM, generally in the light source, they are not as common as more established 'table top' machines such as FTIR, but success has been found using IR-SNOM [60, 97]. There are other high resolution IR imaging techniques available such as tip enhanced raman spectroscopy (TERS). The TERS has a complication though that to achieve high resolution a very intense source light source is needed to produce an adequate signal, which will often result in damaging a biological sample.

5.2 SNOM terminology

An important idea to clarify is that a typical SNOM scan results in six images with two SNOM images, two reference images and two topographical images. There is two of each image type as the SNOM was usually run to operate in both the forward direction and also the backward direction. The two sets would be collected at the same time as for the forward image the SNOM tip would travel from left to right in respect to the images and once it reached the end of the line it would then travel back collecting a measurement at each of the positions previously images, hence building the backward image form right to left. Once the tip had reached the end of the backwards row it would move up ward to scan the next line. This method was chosen as it produces two images in each scan which can be cross compared for possible scanning artefacts or possible even combined to remove noise. But this method also protects the tip from being moved quickly over the sample at the end of the forward scan which may cause damage to the sample and tip if they collide.

This is also a good place to outline various terms which will be utilised repeatedly throughout the chapter. SNOM images or IR images are the images

generated by the measurements taken by the MCT which is situated at the end of the SNOM fibre and so are the image representing the sample. The reference image is complied from measurements taken in sync with the SNOM using a pyro detector to record the intensity of the FEL beam. Although the reference signal is not spatially dependent it is displayed as an image so that it can be compared easily to the SNOM image. The final image type is a topography image which is a recording of the relative height of the SNOM tip. This is also recorded in sync with the SNOM image and gives a insight into the 3D structure of the sample's surface. Finally a data set is defined as a collection of SNOM scans taken in the same area at different wavelengths.

There are also two main types of data to collect when using the SNOM, either a 2D dimensional image which represents the variation in the light intensity spatially as the tip moves across the surface or a spectrum collected over time at a single point of the sample. By gradually varying the wavelength of the probing light over time the spectrum can be used to describe the variation of in the absorbance at a small area of the sample.

5.3 Image preparation

As with the FTIR data sets there is often a degree of pre-processing required to prepare the SNOM data for analysis. This was generally a more significant task for the SNOM data as the images were often noisier due to the low intensity signal collected by the MCT detector at the end of the SNOM fibre. The very weak MCT signal results in a diminished signal-to-noise ratio and therefore required processes to improve the quality the data. A custom program was built which contains the various processes needed and was able to give a real time view of how each stage affected the data, which allowed each image to be processed optimally. As the SNOM is a scanning probe instrument there are also effects found within the SNOM images that are not present in the FTIR data which therefore required specialised corrections. The various stages available to improve the SNOM images

are outlined in detail in the following subsections, which have been arranged in such a way to mirror the order they would be used to correct an image. The pre-processing program was able to run in an automated manor allowing for many images to be corrected quickly, which gave a good overview to which data sets were likely to be promising. Though to achieve the best possible images for the analysis it was operated manually for each image so that each stage could be used optimally to give the best image. The goal of all pre-processing is to use as little as is needed to bring the raw data to the required level of quality and therefore it was common not to use all the stages available as some images didn't require them.

Dropout correction

The FEL was occasionally prone to infrequently producing a very weak pulse of light which would not have the intensity required to be registered by either the MCT or the reference detector. This would result in 'dropout' pixels within the SNOM and reference images as each detector gave a signal close to nought. 'Dropout' pixels would occur 2-3 times on average within a 125 x 125 pixel image and would have sufficient spacing that they were never next to each other within the image. As the pixel value would drop close to zero the normalisation methods described later would not be an adequate solution. The correction would first find any 'dropout' pixels by finding any which had a value less than 10% of the images mean pixel value. Any 'dropout' pixels which were found would then be replaced by the average of its nearest surrounding pixels. For this correction not to cause complications in the later normalisation stage it had to be made in both the SNOM and reference images.

Normalisation

Normalisation is a standard correction method common in many spectroscopy techniques, which aims to remove any effects due to the variation in the light

source's intensity in the collected data. This is possible in the SNOM instrument as it is equipped with a detector which records a reference signal measuring the variation in the intensity of the FEL beam at the SNOM table. The SNOM records the reference signal every time it takes a measurement of the SNOM signal and therefore every pixel in the SNOM image has an associated value of the FEL beam's intensity. Normally each reading in the experimental signal (SNOM signal) would be divided by the appropriate value in the reference signal pixel for pixel, which if ideal would remove any variations in the SNOM image due to the FEL beam's power fluctuating throughout the scan. This is not the case for the SNOM though, as the SNOM signal and the reference signal are not entirely equatable. This is due to the reference signal measuring the entire beam's intensity, whereas the SNOM signal is collected from a very small area of the focused beam. The SNOM signal is therefore sensitive to the internal structure of the FEL beam and any positional movements of the beam. This means the two signals can't be simply divided on a pixel to pixel basis. Although sometimes this type of normalisation was found to improve SNOM images it was also likely to add extra noise due to the discontinuity between the two signals.

A more general approach was therefore needed to be able to use the reference signal to normalise the SNOM image. This was done by plotting each pixel in the image on a graph of the reference signal against the SNOM signal. If there was variation in the SNOM image due to the fluctuation of the FEL beam's intensity, the graph would result in a linear correlation as an increase in the FEL beam's intensity would result in an increase in the SNOM signal and vice versa, Figure 5.1 shows an example of such a graph.

A linear line can then be fitted to the distribution to give a general relation between the SNOM signal and the reference signal. The fit could be used to normalise the SNOM image as each pixel would be divided by the value from the fit at it's associated reference value rather than simply the reference value itself. Within the preprocessing program the fit could be generated automatically using the inbuilt 'fit' function in MATLAB or it could be formed manually and manipulated in real time to optimise the normalisation of the SNOM image. This

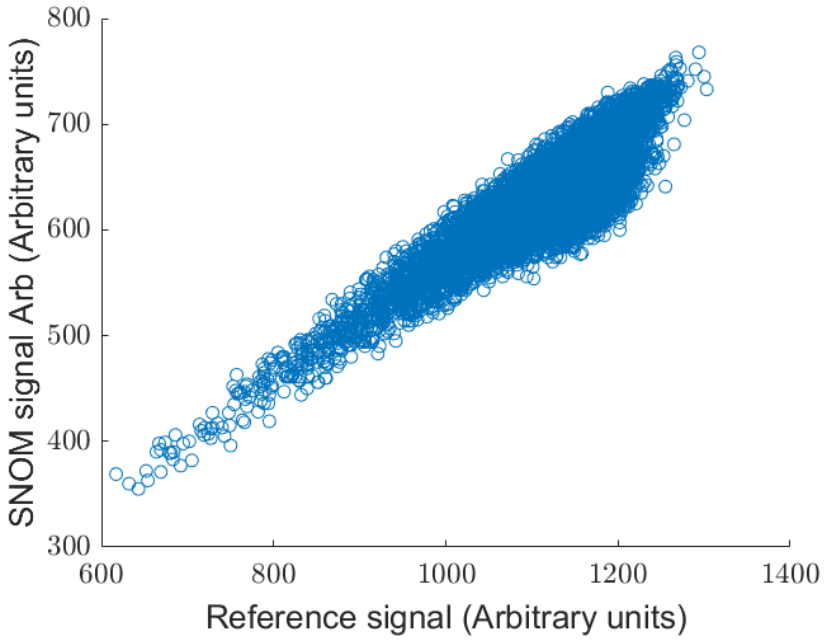

FIGURE 5.1: A graph showing the correlation between the SNOM signal and the reference signal taken from a SNOM scan of a OE 33 cell.

was found to be one of the most powerful processes to extract information from images which appeared not to have a large amount of detail in the raw data, as Figure 5.2 demonstrates.

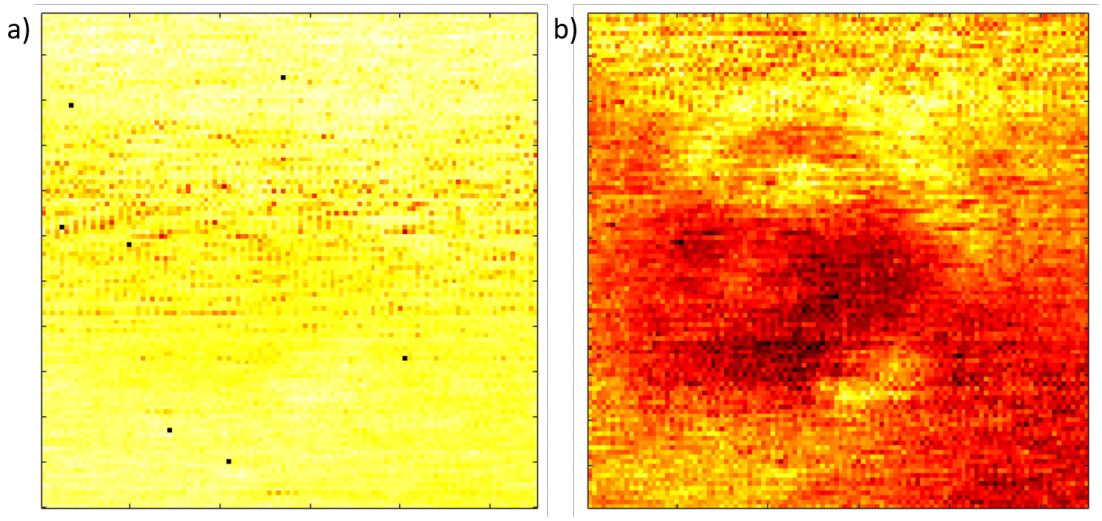

FIGURE 5.2: Figure showing the result of using the dropout and normalisation corrections. a) Shows the raw SNOM image taken of a OE 33 cell and b) Shows the result of using the correction on the raw image.

Piezoelectric stage correction

A common problem encountered when using piezoelectric stages is the nonlinear movements due to the stage not moving proportionally to the driving signal which is used to control it. This issue was found to exist in the SNOM sample stage which meant the images produced were prone to artificial stretching due to the varying distances between each measurement. This isn't an issue if the analysis being carried out on the data is not spatially dependent, but if the SNOM images are to be compared back to the original sample the non-linearity within the piezoelectric drive needs to be accounted for.

The first step is to characterise the nonlinear movement of the stage so that a tailored correction can be made. To do this a calibration sample consisting of a silicon base with regular sized and spaced gold squares placed on top was imaged using a wire with a very sharp point. A wire was used instead of the SNOM fibre as it could be made into a much sharper point and as this study only requires the topography it was ideal. Figure 5.3 shows an optical image of the calibration slide and the raw topographical images collected by the SNOM in both the forward (collected left to right) and backward (collected right to left) directions.

The forward and backward images in Figure 5.3, show that there is also an asymmetry between the two directions as the most significant stretching occurs at the beginning of each line scan, which is on the left side for the forward scans and the right side for the backward scans. This means an individual correction has to be found for both the forward and backward scans. The sizes and spacing of the gold squares had been measured using a calibrated optical microscope and so the distance from one edge to another could be calculated. A characterisation curve could therefore be made by plotting the real spacial distances of each edge from a specified edge against the observed distance depicted within the SNOM images. Figure 5.4 shows the calibration curves for both directions, with a fitted third order polynomial which describes the relationship between the perceived tip position and it's real position.

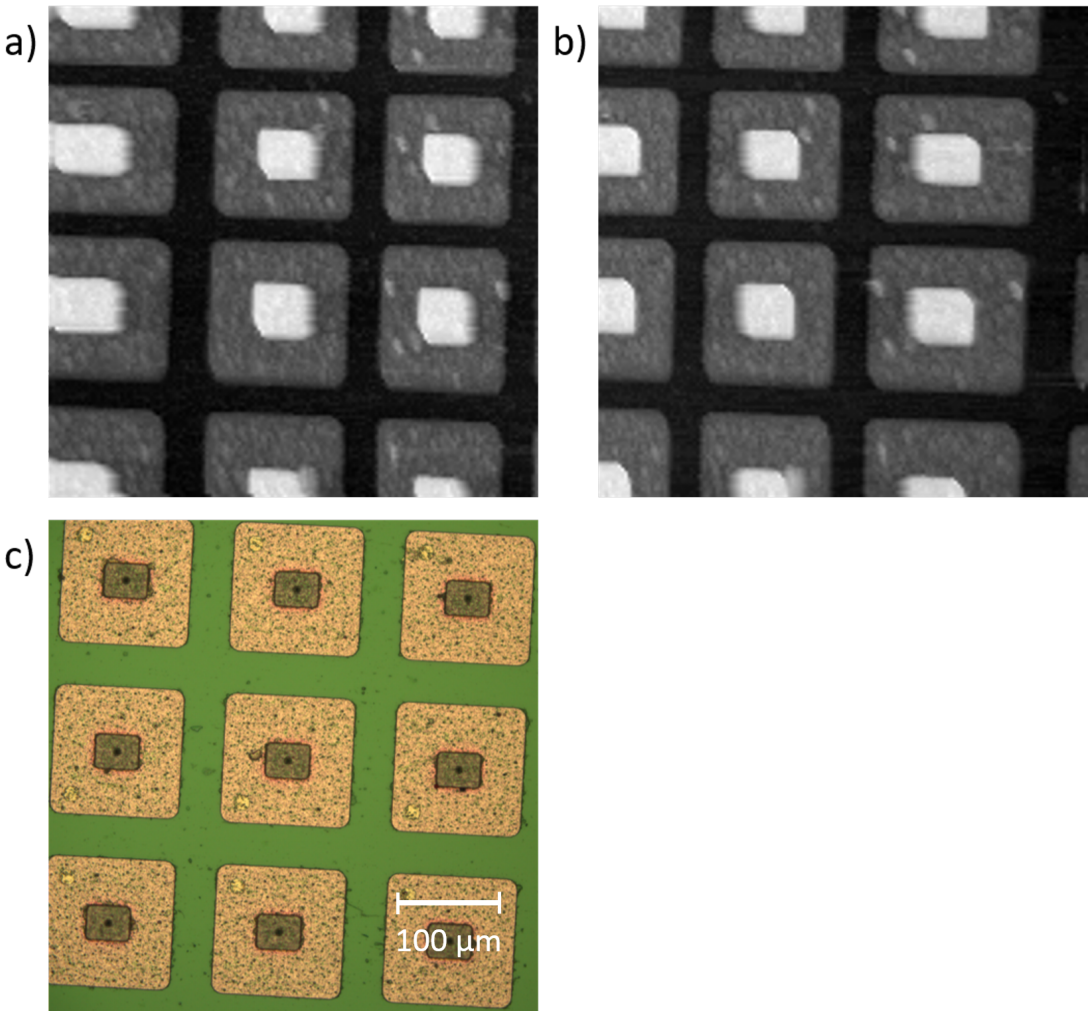

FIGURE 5.3: Highlighting the stretching of the SNOM image due to the non-linear behaviour of the piezoelectric stage. a) Topography image of the calibration sample taken in the forward direction, b) the same scan but in the backwards direction and c) optical microscope image taken of the calibration grid.

The calibration curves therefore allow for the real spatial position of every pixel to be calculated by finding the relevant point on the curve given it's observed spatial position in the raw image. Each line within the SNOM image could then be interpolated with a new spacing in such away that the resultant image had a uniform pixel spacing of a size denoted before the scan. Figure 5.5 shows the result of using the correction on the calibration images shown in Figure 5.3.

It is clear that post correction the topography image is now much closer to the sample's true topography and by testing multiple full range calibration images it was found to be a reliable method of correction. Smaller images were taken within

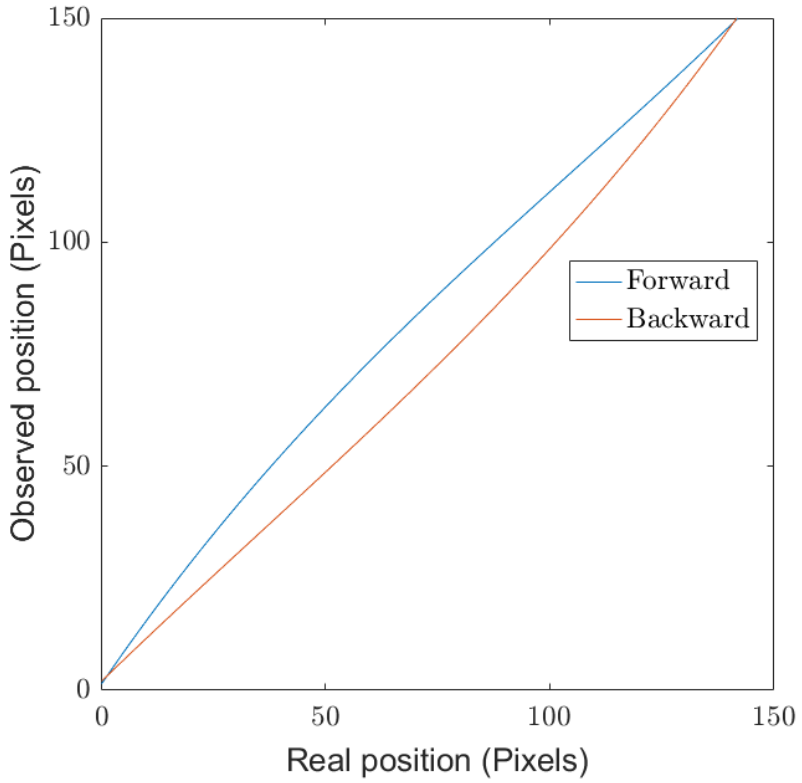

FIGURE 5.4: A graph showing the relationship between the observed and actual potion of the SNOM tip throughout a forward and backward scan.

the full range of the stage and corrected by taking the relevant portion of the calibration curve. When the small images were corrected they were much closer the real topography but were found not to perform optimally as the full sized scan. This must be the result of a relevant portion from the 500 µm calibration curve not properly describing the non-linearity found within the smaller images. The general correction would therefore be used to bring all the raw images close to the real topographical structures, while a separate program was developed for the smaller images which allowed for a user to make small alterations to the calibration curve resulting in the best images. Calibrated optical images or AFM scans were often used to compare the correction's output to assure that the distortion had been properly corrected.

This correction would be applied to both the topographic images and the SNOM images so that the absorption features could be correlated spatially to potential features within the sample. The piezoelectric correction is carried out after the

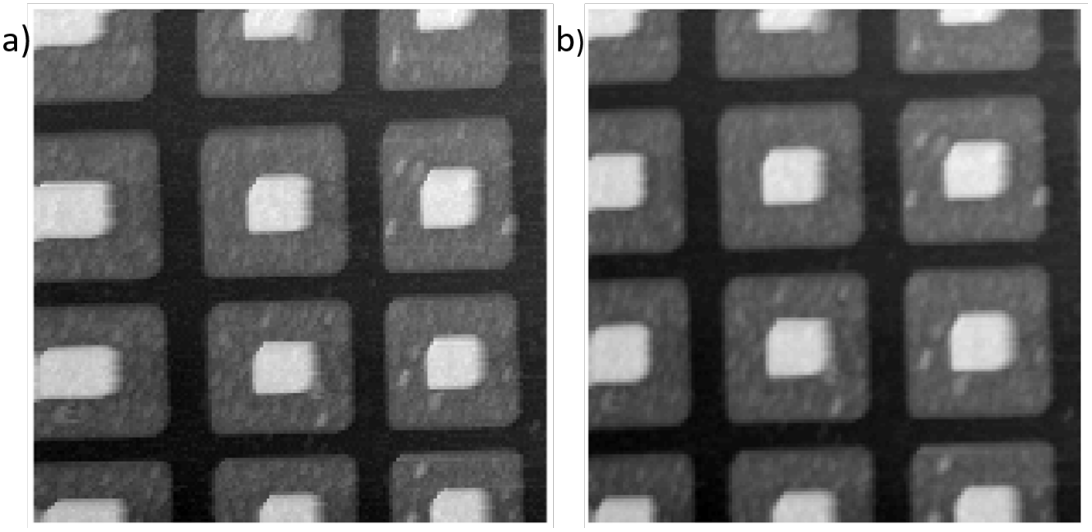

FIGURE 5.5: Figure showing a) A raw topography image taken by the SNOM of a calibration grid, b) Is the same topography after being corrected.

normalisation stage as interpolating the reference image which may mostly contain variations due to noise isn't ideal and may detrimentally affect the performance of the normalisation program.

Image registry

To properly compare the images within a SNOM data set, it is important to be confident that all the images are aligned and are therefore displaying the same area of sample. It was found that between scans the SNOM tip was liable to drift by $< 0.5\,\mu m$, due to sample movements or thermal effects. This is a negligible effect for large images where the pixel sizes are much greater than the drifts, but for the much smaller high resolution images where the pixel size is sub-micron, the drift becomes visible and would have a serious effect on the analysis if not corrected. To produce data cubes like the FTIR it is crucial that the SNOM data is properly aligned so that the pixels within each image are correctly positioned.

The images taken within a SNOM set therefore need to be co-registered if reliable comparisons are going to be made between them. Using the SNOM images would be unreliable for registry as each image may vary greatly due to the interaction of the light with the sample, but the topographical image produced in every

scan are ideal as they show the scan area and should be identical for each scan within the set. To align the images, each topographical image is compared to the topography of the first scan and the program would move one topography image against the other over reasonable distances and measure the average difference between the commonly shared area, which when properly aligned would be at a minimum. Once the offset for each image within the set was found the images, including the SNOM and reference images would be cropped to the common area shared by all the scans.

Median filtering

Filtering was often avoided as it can overly blur the SNOM images removing detail but median filtering would sometimes be used if the noise within the image was still too large even after normalisation. Median filtering is a common noise reducing image analysis process and works by replacing each pixel with the median of its nearest surrounding pixels, hence removing the more extreme values. This may help removes noise but was always used manually so that the resultant image could be assessed to see if the filtering was beneficial.

5.4 High resolution study

As the main aim of using the SNOM was to assess its ability to generate diffraction limit breaking resolutions in images using the IR-FEL light source which was incorporated into the ALICE accelerator. Such images would not be possible with traditional far-field instruments, such as FTIR as they are fundamentally limited by the diffraction limit. To achieve such high resolutions an etched tip with an aperture diameter of $\approx 0.1\,\mu m$ was rastered over the samples surface. For the images shown in this study the SNOM was operated in reflection mode, which means the FEL light strikes the sample at a grazing angle and a small amount of reflected light is captured by the SNOM tip. As the tip captures the reflected light the SNOM images are only sensitive to the chemistry near the sample surface.

To test the ability of the SNOM instrument to produce submicron resolution images, five scans were taken of a small 5 µm by 5.5 µm area within a single OE 33 (oesophageal adenocarcinoma) cell. Each scan used a different wavelength for the probing light (5.71, 6.06, 6.5, 7.3 and 8.05 µm), so that chemical sensitivity of the SNOM instrument could be assessed. The wavelengths were carefully chosen to highlight the presence of important biomarkers which are common to many biological structures, apart from the 7.3 µm wavelength, which was chosen as it isn't associated to any key biological absorption bands and therefore should act as a control. Table 5.1 shows which biomarkers were targeted by which wavelengths.

Wavelength µm	Targeted biomarkers
8.05	DNA/RNA
7.3	Control
6.5	Amide II (β-sheet)
6.06	Amide I (α-helix)
5.71	Lipids

TABLE 5.1: A table showing which biomarkers were targeted by each wavelength.

Figure 5.6 shows a optical microscope image taken of the OE 33 cell used within this study, along with the SNOM topography image of the same cell. The black rectangle on the topography image shows the scam area for the images within this high resolution study.

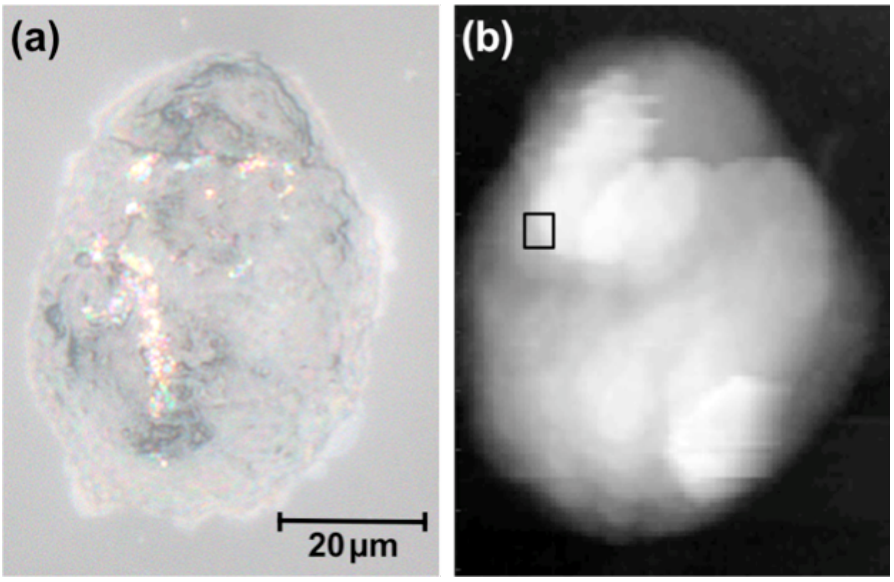

FIGURE 5.6: Various images of the same OE 33 cell a) optical microscope image of the OE 33 cell used within this high resolution study, b) SNOM topography taken of the same cell, with a black rectangle highlighting the area scanned in the high resolution images.

The high resolution SNOM images taken within the highlighted area are shown in Figure 5.6. Each of the scans have been registered together using the topography images and cropped to guarantee that all SNOM images shown in Figure 5.7 represent the same area and are not offset. Within the SNOM images areas which had an intense light signal are represented as white while areas with low light signals are coloured red.

It is clear from the SNOM images that depending on the wavelength of the probing light, different features are present within the images, for example the 8.05 µm image shows a very intense streak with areas of low intensity on either side. The 6.06 µm, 6.5 µm and 7.3 µm images instead generally mirror the topography, showing the entire top right corner of the images having an intense signal, unlike the 8.05 µm image. The 5.71 µm image is the only image which doesn't show clear contrast at the topography features edge. This is due to the pulse to pulse noise

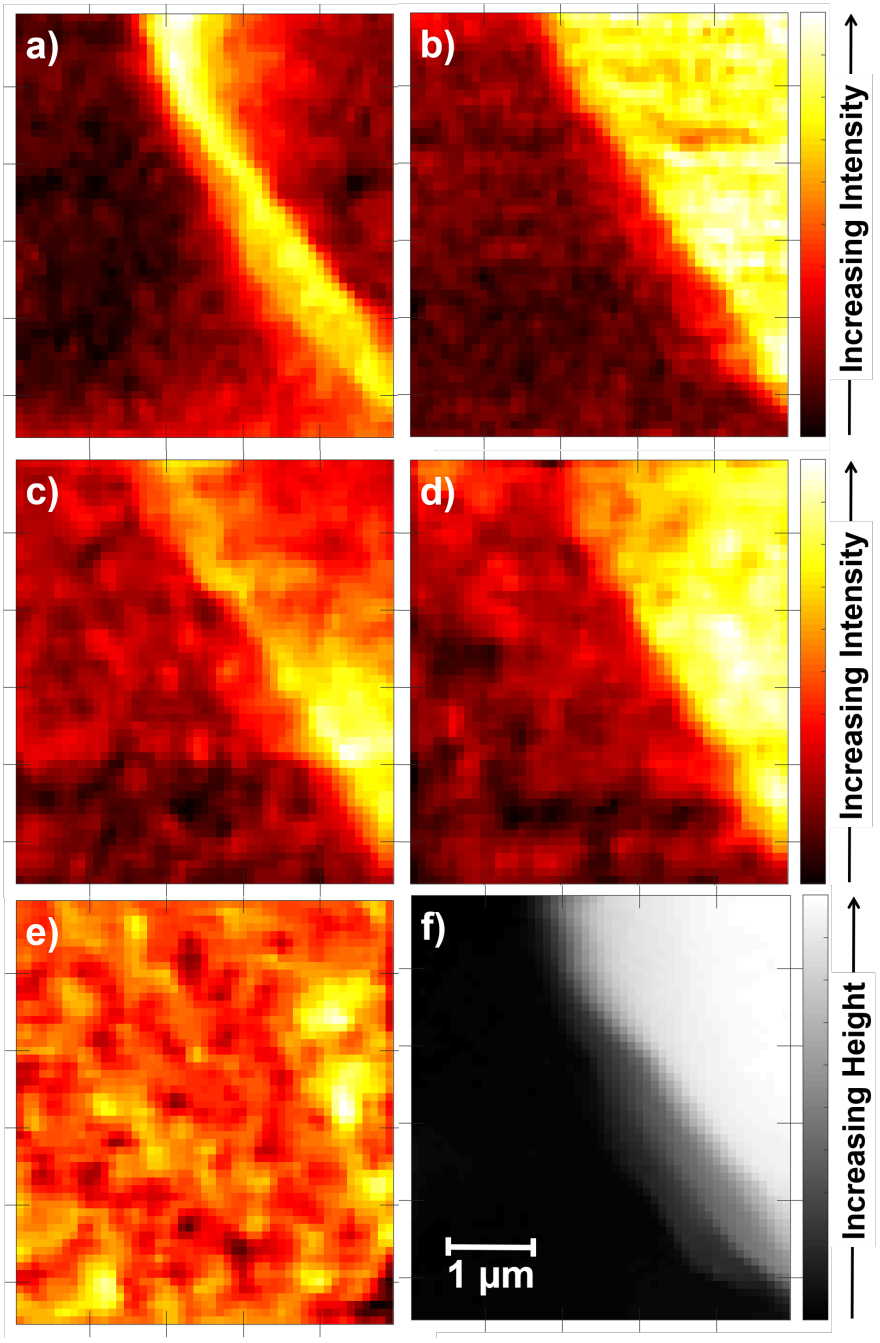

FIGURE 5.7: Five high resolution SNOM images taken at various wavelengths and a topographical image. a) - e) Are the SNOM images using $\lambda = 8.05$, 7.3, 6.5, 6.06 and 5.71 µm respectively and f) Is the topography image taken from the same area.

being considerably larger when the FEL producing 5.71 µm light as it is lasing near it's limits and can become unstable. The increased noise coupled with the very weak lipid signal due to it being only generated by a very thin lipid bilayer means that the 5.71 µm image doesn't appear to contain any significant contrast.

It is possible that the tip and/or sample has changed between the subsequent scans and this is the cause of the varying image contrast. To assess if either sample or tip changes caused the varying contrast, the first scan (8.05 µm image) was repeated after all the other images had been collected. The original image and the SNOM image from the repeated scan are compared in Figure 5.8.

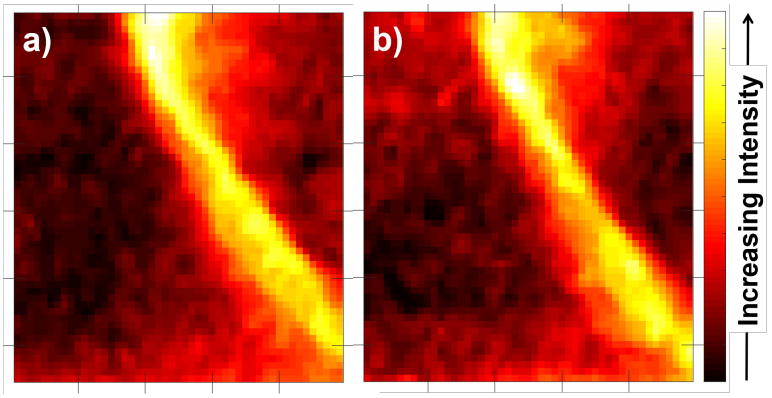

FIGURE 5.8: Comparison of the original 8.05 µm image and a repeat scan a) The original 8.05 µm image, b) The SNOM image from the repeated scan.

It is clear that the repeat scan is very similar to the original image and therefore the tip and sample are not the cause for the change in the features within the SNOM image. The variations are also unlikely to be solely due to topographical artefacts as all the topography images are consistent with each other, but the SNOM images show considerable variations in contrast. To properly assess the spatial positions of the features seen in the SNOM images, line profiles were taken from the same position on each of the SNOM image. The line on the topography image shown in Figure 5.9 highlights the region the line profiles were taken from. The line profiles taken from the SNOM images are shown in Figure 5.10.

What is clear from the line profiles is that the leading edges of the features shown in the SNOM images occur at different spatial position. The leading edge of the feature shown in the 8.05 µm images clearly start before any of the features shown in the other SNOM images. The line profiles also show that even though the 6.06, 6.5 and 7.3 µm images appear similar the edge starts and rises differently within each image. For example the 6.5 µm edge starts before the 7.3 µm edge and rises much steeper than the 6.06 µm edge. These observations demonstrate the each of the SNOM images show fine detail differences which distinguish them from

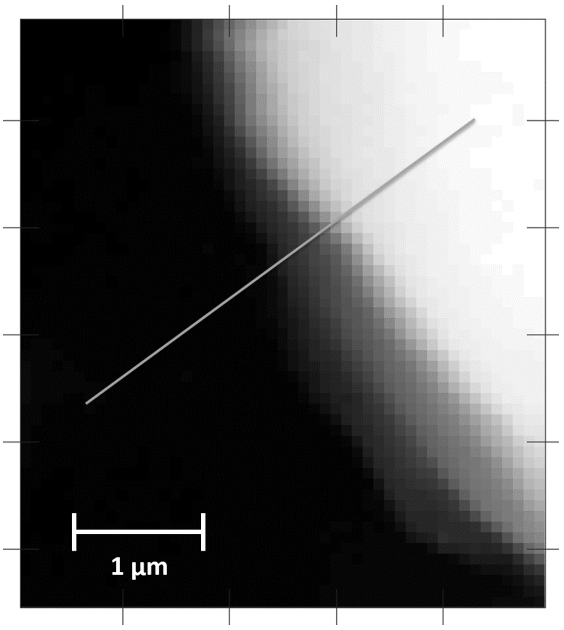

FIGURE 5.9: A topography image with the region used to generate the line profiles highlighted.

each other, again implying that the features shown are not simply topographical effects.

As the true size of the fibre aperture diameter is unknown the resolution has to be estimated using the features visible within the images. The resolution within an image can be estimated by measuring the distance required for the intensity to increase from 20% to 80% over a step feature. By using this rule and the 6.5 µm line profile an estimation of $\approx 0.15\,\mu m \pm 0.05\,\mu m$ can be found. By using this value of the resolution the SNOM instrument can be shown to achieve a resolution which is over 20 times that of optimal resolution determined by the far-field diffraction limit.

To further assess the structure found in the 8.05 µm SNOM images, multiple overlapping scans were taken so that they could be combined to form a mosaic image, the result of which is shown in Figure 5.11. The mosaic shows that the feature has a clear beginning and end, with finite dimensions of $\approx 11\,\mu m$ by 0.7 µm. The fact that this object has a finite size and does not continue to follow the topography, which is also shown in Figure 5.11, provides more evidence that the features seen within the SNOM images are based on the chemistry of the sample

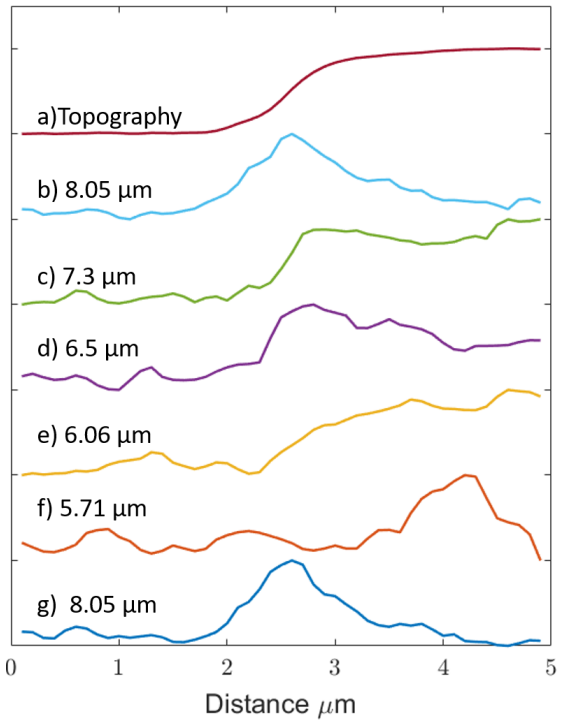

FIGURE 5.10: The line profiles taken from each of the SNOM images within the resolution study.

and are not just simply image artefacts. These dimensions are also consistent with the size of human chromosomes, which while can not be definitively proven in this case, shows the potential of a high resolution technique such as SNOM.

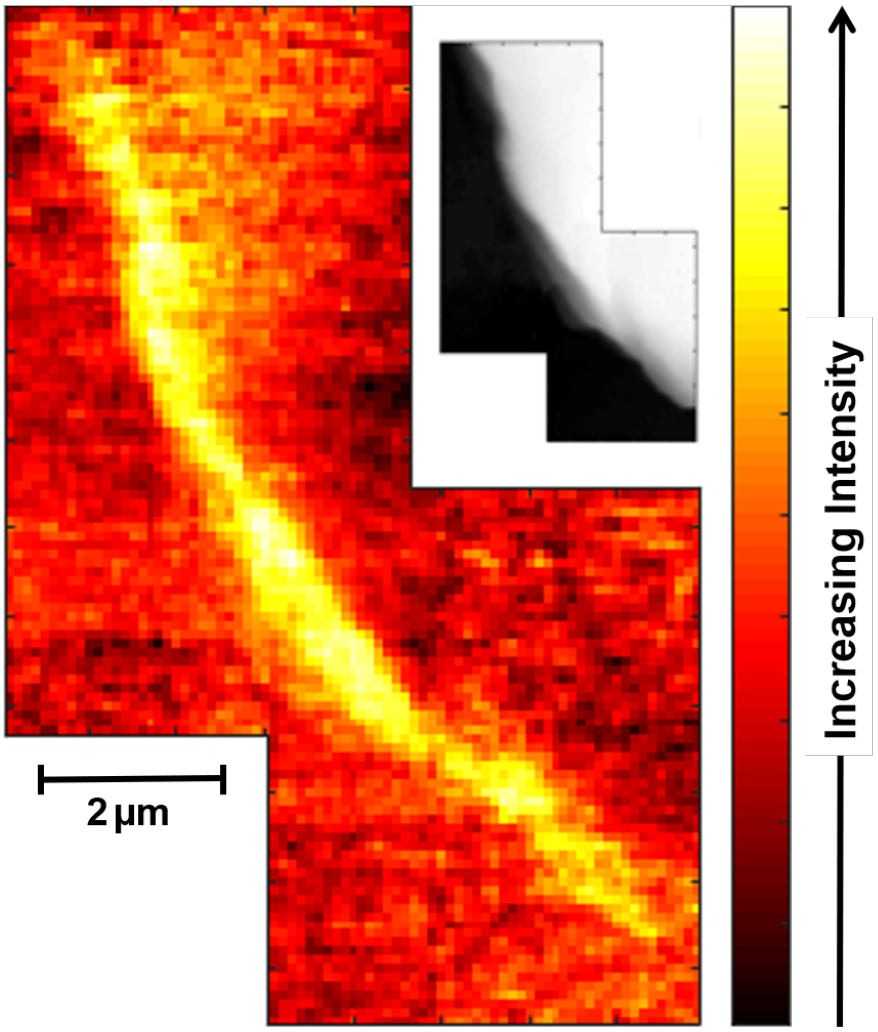

FIGURE 5.11: A mosaic made of three individual SNOM images which have been combined to show that the structure seen within the 8.05 µm SNOM images has a finite length

5.5 Evaluation of aperture SNOM for biomedical applications

To contrast and compare both the reflection and transmission SNOM imaging along with FTIR spectroscopy, all 3 techniques were used to image the same OE 33 cell shown previously. By comparing the images it might be possible to deduce any potential differences between them, giving a better indication as to their strength and weakness.

The same cell shown in Figure 5.6 was imaged using the same FTIR used in

Chapter 4 but with a high magnification condenser. This reduced the image size to 140 µm x 140 µm, with a pixel size of 1.1 µm. It is important to note that although the higher magnification condenser can shrink the pixel size, the instrument is still restricted by the diffraction limit. Figure 5.12 shows all the SNOM images along with the appropriate slices taken from the FTIR data cube at the same wavelengths as used with the SNOM.

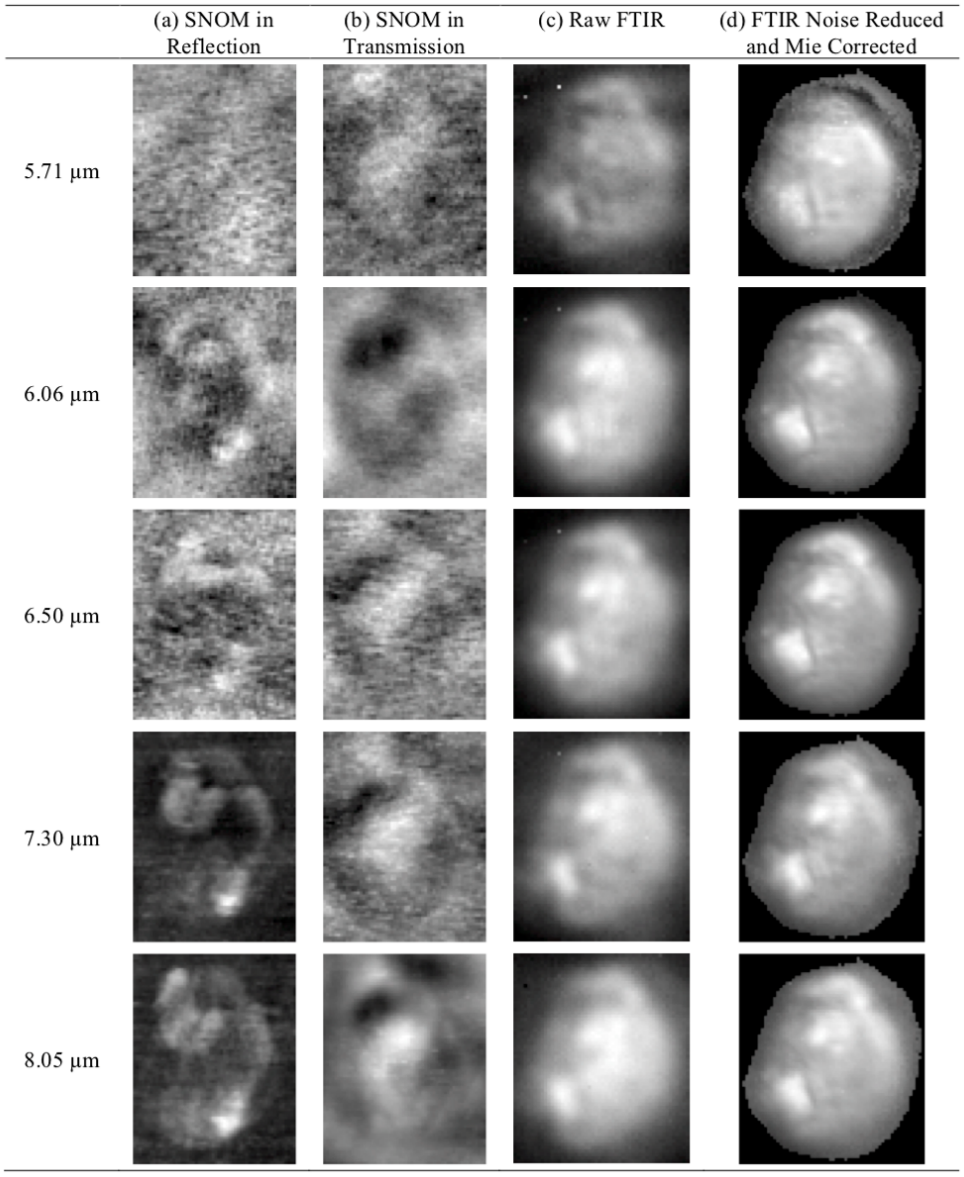

FIGURE 5.12: Comparison of SNOM and FTIR images of an OE33 cell. (a) SNOM in reflection, (b) SNOM in transmission, (c) raw FTIR images and (d) FTIR images after noise reduction and Mie scattering correction. The SNOM images are 58 µm x 75 µm while the FTIR images are 66 µm x 75 µm.

The reflection and transmission SNOM images were taken one after the other during an experimental run at ALICE, the FTIR data was taken 7 months after the SNOM images. To assess if the cell had changed between the SNOM and FTIR scans the cell was imaged using a atomic force microscope (AFM), which is capable of very high spatial resolution topography images. Figure 5.13 shows both the AFM and SNOM topographies.

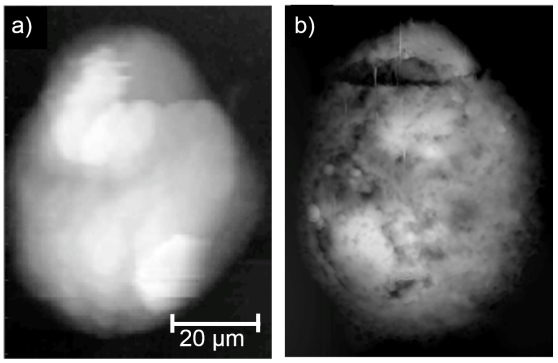

FIGURE 5.13: Comparison of a) a SNOM topography and b) an AFM topography of the same OE 33 cell taken 7 months apart.

The AFM image obviously has a much higher spatial resolution (0.07 µm) compared to the SNOM which is to be expected. Figure 5.13 clearly shows that there has been a considerable change over the 7 months between both images being taken. Although the cell is chemically fixed to stop it decaying it does not prevent the cell from 'settling' while it was stored. As the structures seen in both images are different it implies that the internal contents of the cell will have moved. This means that the SNOM and FTIR images cannot be compared directly as the distribution of the chemicals within the cell has changed.

Although direct comparisons cannot be made between the SNOM and FTIR images, they can be studied to assess the performance of the SNOM at gaining higher spatial resolution images. Figure 5.14 shows the 8.05 µm SNOM image and 8.05 µm FTIR image, where two line profiles have been taken in approximately the same positions in both images. It is important to note again that it is not expected that the general structures will be the same as the cell is known to have changed.

Clearly the SNOM image shows a much greater magnitude of signal variation and higher resolution spatial features compared to the line profiles from the FTIR image. This is to be expected as the FTIR is still restricted by the diffraction limit and therefore the line profiles are essentially smoothed as the each point is averaged with it's neighbours, because the pixel size is smaller than the diffraction limit.

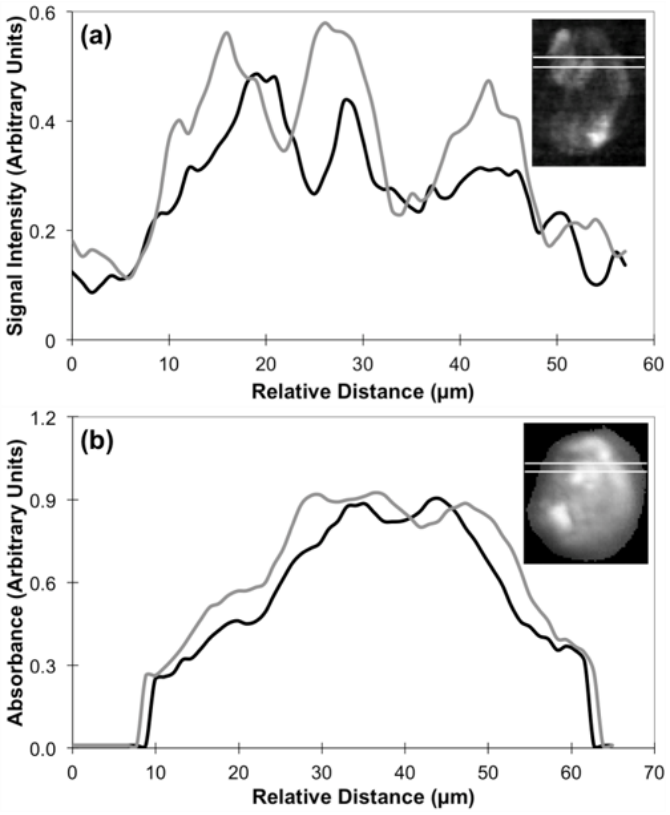

FIGURE 5.14: Line profiles through a OE 33 cell at 8.05 µm. (a) Line profiles through the SNOM in reflection image and (b) line profiles through the noise reduced and Mie corrected FTIR image. The inserts show the locations of the line profiles. The black line profiles are taken from the upper line and the grey line profiles are taken from the lower line through the small insert images.

As the SNOM images were taken around the same time they can be compared to assess if the reflection and transmission images contain the same information. Figure 5.15 shows the topography, reflection and transmission SNOM images at 8.05 µm. Visually it is clear that the reflection and transmission images appear to have very little correlation, as they share very few common features.

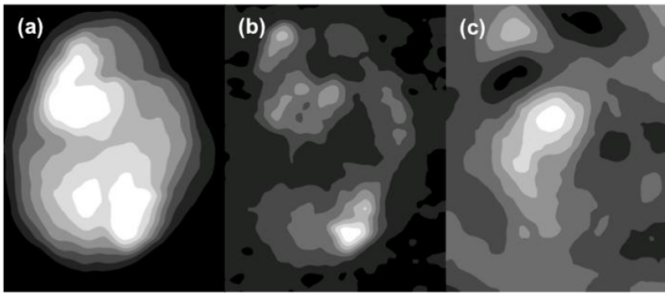

FIGURE 5.15: Contour Plots showing (a) SNOM in reflection topography, (b) SNOM in reflection at 8.05 µm and (c) SNOM in transmission at 8.05 µm.

To quantify the degree of correlation between these images the correlation coefficient was calculated for each pairing. A correlation coefficient for any two images can be calculated by the sum, over all the pixels within the cell area, of the products of the pixel values in the two images which have been offset from the mean value and normalised by the variance in each image. A correlation coefficient of 1 implies that the two images are identical, -1 indicates that they are perfectly anti-correlated and 0 means there is no correlation between the images. The correlation coefficients of the reflection, transmission and topography images are shown in Table 5.2.

Image pair	Correlation coefficient
Reflection-Topography	0.43
Transmission-Topography	0.31
Reflection-Transmission	0.01

TABLE 5.2: A table showing the correlation coefficients between the reflection, transmission and topography SNOM images of an OE 33 cell.

The values shown in Table 5.2 indicate there is a degree of correlation between the SNOM IR images (reflection and transmission images) and the topography, but the value is low enough to imply that the features within the IR images are not solely due to the topography. As expected there is very little correlation between the reflection and transmission images, with a correlation coefficient of 0.01.

The correlation between two images can be shown visually with correlation plots as demonstrated in Figure 5.16, which show the comparison of the reflection and transmission images at all wavelengths along with the topography. The areas of the cell where the reflection and transmission images correlate for a given wavelength are coloured green, the areas of anti-correlation are coloured red and the areas of no correlation are grey. The correlation coefficient for the cells as a whole is displayed beneath each image. As the topography image demonstrates a very high correlation of 0.97, it implies that the cell was not significantly damaged or structurally altered during the reflection or transmission scans.

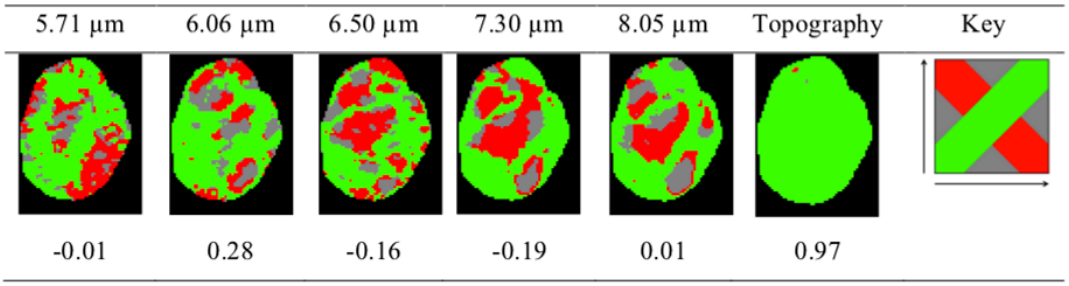

FIGURE 5.16: Correlation plots and coefficients for SNOM in reflection vs SNOM in transmission. The correlation plots show the correlation spatially between the two images for a given wavelength. Green indicates that the two images correlate well, red indicates that the two images are anti-correlated and grey implies there is no correlation.

As the correlation coefficients at all the wavelengths are low it indicates that there tends to be very little correlation between the reflection or transmission scans. The fact that the IR images are not correlated but the topography images are demonstrates that reflection and transmission are sensitive to the chemical composition of the cell. The lack of correlation in the IR images also implies that the two techniques have different sensitivities which is mostly likely sample thickness.

The same method for correlation used to compare the reflection and transmission sets can be applied to correlate the SNOM images at different wavelengths. Table 5.3 shows the correlation coefficients of all the comparisons of IR images and topography within the reflection set.

	5.71 µm	6.06 µm	6.50 µm	7.30 µm	8.05 µm	Topography
5.71 µm	1.00	0.08	0.03	-0.01	-0.05	-0.03
6.06 µm	0.08	1.00	0.52	0.24	0.10	-0.40
6.50 µm	0.03	0.52	1.00	0.57	0.39	-0.08
7.30 µm	-0.01	0.24	0.57	1.00	0.75	0.38
8.05 µm	-0.05	0.10	0.39	0.75	1.00	0.43
Topography	-0.03	-0.40	-0.08	0.38	0.43	1.00

TABLE 5.3: table showing the correlation coefficients for the reflection SNOM images.

The first relation that can be found in Table 5.3 is the correlation of the IR images with the topography. Both the 8.05 µm and 7.3 µm images were somewhat correlated with the topography, 6.5 µm and 5.71 µm showed little correlation and 6.06 µm displayed a degree of anti correlation. None of the reflection IR images therefore demonstrate that they are entirely dependent on the topography. A strong correlation is demonstrated between the 8.05 µm and 7.3 µm with a coefficient of 0.75.

There is a strong correlation of 0.52 between 6.06 µm and 6.5 µm, which are attributed to α-helices and β-sheets respectively. This is understandable as both are components of the secondary structure of proteins. The 5.71 µm image showed very poor correlation with all the images, which can be attributed to the poor quality of the 5.71 µm image. As shown in the high resolution study 5.71 µm is at the edge of the ALICE-FEL operating range, so the signal to noise can often be poor.

A similar table can be produced for the transmission SNOM set, shown in Table 5.4.

	5.71 µm	6.06 µm	6.50 µm	7.30 µm	8.05 µm	Topography
5.71 µm	1.00	0.11	0.70	0.74	0.66	0.25
6.06 µm	0.11	1.00	0.39	0.22	0.20	-0.43
6.50 µm	0.70	0.39	1.00	0.70	0.74	0.08
7.30 µm	0.74	0.22	0.70	1.00	0.74	0.26
8.05 µm	0.66	0.20	0.74	0.74	1.00	0.31
Topography	0.25	-0.43	0.08	0.26	0.31	1.00

TABLE 5.4: Table showing the pixel correlation coefficients for the transmission SNOM images.

Table 5.4 shows largely the same relations that were displayed in the reflection images. The highest correlation is still between the 8.05 µm and 7.3 µm images with a coefficient of 0.74. The IR images also have the same relationship with the topography as shown in the reflection dataset, apart from 5.71 µm that has

a slightly stronger correlation with the topography. This is most likely due to the quality of the 5.71 µm transmission image being improved compared to it's reflection counterpart. This improvement may also be the reason as to why the correlation between 5.71 µm and 8.05, 7.3 and 6.5 µm have also increased significantly from the reflection images. The correlation coefficient of both the 6.06 µm (α-helices) and 6.5 µm (β-sheets) doubled when compared to the 8.05 µm image.

The same study can be carried out on the corrected FTIR data, which is shown in Table 5.5. The results within Table 5.5 show that the FITR images are massively correlated to the topography which was taken using the AFM, with correlation coefficients ranging from 0.93-1. This is not to say that there isn't chemical information present in spectra it is just simply dominated by the topography of the sample. This highlights the benefit of the MA algorithm processing ratio values as theY are thickness independent.

	5.71 µm	6.06 µm	6.5 µm	7.3 µm	8.05 µm	Topography
5.71 µm	1.00	0.96	0.94	0.94	0.94	0.93
6.06 µm	0.96	1.00	0.97	0.97	0.97	0.97
6.5 µm	0.94	0.97	1.00	1.00	1.00	1.00
7.3 µm	0.94	0.97	1.00	1.00	1.00	1.00
8.05 µm	0.94	0.97	1.00	1.00	1.00	1.00
Topography	0.93	0.97	1.00	1.00	1.00	1.00

TABLE 5.5: A table showing the pixel correlation coefficients for the FTIR images which have been noise reduced and Mie corrected .

By studying the coefficients in Table 5.3, 5.4 and 5.5, it is clear that the SNOM images are less dominated by the topography compared to the FTIR images. It is also clear that the reflection and transmission images tend to show different features, which is demonstrated in Figure 5.16. As the cell is the same for both sets this implies that reflection and transmission are imaging different thickness of the sample.

5.6 SNOM experiment conclusion

The high resolution study of the OE 33 cell demonstrates the true power of the SNOM, which is to produce diffraction breaking spatial resolutions in chemically sensitive images. IR-SNOM for biological studies has many challenges including acquiring an IR light source which is both stable, powerful and is also practical enough to meet the needs of a SNOM instrument. The IR transmitting fibres are delicate and are often hard to reliably prepare into an etched tip with an aperture size capable of high resolution imaging. The data acquisition rate is very poor compared to many other techniques such as FTIR, when the SNOM is coupled to a FEL as it is locked to the pulse rate of accelerator (10Hz).

But given these complexities the results shown within this chapter demonstrate that a SNOM is capable of producing images with sufficiently high resolution that it would be impossible to recreate with far-field IR techniques. As SNOM is still somewhat in its infancy as a technique the field is open to major advancements which may push it's abilities and applications into new areas. One potential application for SNOM is the imaging of the internal chemical distributions with in diseased cells at various stages of development. It is also applicable for experiments where the spatial resolution is critical, such as studies involving very thin membranes, which are impossible to see using far-field techniques.

By imaging the same cell in both reflection and transmission, they were able to be compared demonstrating that they appear to be sensitive at different depth of the sample. The SNOM images could then be compared to the FTIR image of the same cell which showed that the SNOM is much less sensitive to the topography than the FTIR.

The objective for the SNOM experiments was to assess the SNOM's potential for medical studies and demonstrate it's key strengths. The SNOM images shown in Figure 5.7 are the same size as a single pixel from a standard FTIR image. Within these images there is clear detailed contrast showing the image is related to the chemical structure within the cell. This contrast was shown to vary depending

on the wavelength of the probing light, showing that the SNOM image was sensitive to the wavenumber. The resolution was estimated to be $0.15\,\mu m \pm 0.05\,\mu m$, which beats the diffraction limit by over a factor of 20. This clearly demonstrates that the key strength of the SNOM which is to produce images of very high spatial resolution.

Chapter 6

Conclusion

6.1 Conclusion

This thesis demonstrates that by studying oesophageal samples using infrared (IR) instruments there is a critical amount of chemical information that can be gained. By using powerful techniques such as FTIR, large amounts of spectral information can be collected very quickly. The samples can be either tissue biopsies or cells which mean such an instrument is very flexible to the needs of the user. Due to the size of the data generated by FTIR an automated learning process is needed to extract the key information which would allow for a classification model capable of distinguishing between the various samples. Metric analysis (MA) was developed since a tool was needed which could both act as a 'blackbox' tissue classifier but also a tool which can be used to probe the very subtle trends between the samples. By achieving excellent success rates for correctly labelling spectra the MA algorithm was able to operate as a high performing classifier. This is demonstrated as the MA was able to distinguish between cancer associated myofibroblast and non-cancer associated myofibroblast, which is currently very hard to do with standard diagnostic techniques. Various visual plots were used to show the outputs of MA in an intuitive manor allowing for the easy assessment as to which wavenumbers are important for discrimination. This meets the needs

of a researcher who desires a greater insight into the mechanics within cancer. For the MA to become established it first has to be compared to other forms of commonly used ML algorithms, which in this case was random forest (RF). MA was found to not only outperform the RF at sample prediction but the learning process was also considerably quicker. By using a statistical based ML algorithm the subjectivity which is fundamental to current histopathology is removed. The work demonstrated with the FTIR shows a possible direction that may in the future become a core to histopathology, once various community driven issues such as sample preparation and sufficiently large medical trial studies are conducted.

High resolution imaging techniques are often riddled with complexities and SNOM is no exception to the rule. Throughout my PhD continual upgrades have been applied to the SNOM with each new iteration learning from the lessons of the past. This means that current iteration is a much more reliable and powerful instrument than what the previous versions were. A resolution of 0.15 µm ± 0.05 µm was demonstrated which equates to a resolution that is 20 times better than diffraction limit. Such a high resolution opens new possibilities for IR spectroscopy to be used in medical studies which were previously impossible with far-field instruments. SNOM images taken in both reflection and transmission mode were compared and showed that each seems to have a different sample depth sensitivity, which may be possible to exploit in the future when more understood. A biological structure was imaged within the cell at very high resolution, which is an example of the power of the SNOM.

6.2 Future work

The results of the work on both the FTIR and the SNOM have been promising with MA appearing to be a potentially powerful tool for diagnostics and characterisation, while the SNOM was able to achieve spatial resolutions impossible to most optical instruments.

There is always more that could be done though to further advance the research. The main limitation within the FTIR study was the limited amount of data available. Although the MA has been able to make significant comments as to the important wavenumbers needed for the discrimination of oesophageal samples, the sample size is not large enough for them to be of interest to the medical community in general. Ideally the MA would be tested with a wide range of samples that contained different tissues and cancers while also coming from a wide variety of patients. Samples from a large number of patients should be used within the learning process to ensure that the classifier model is robust as possible. A more detailed study with samples of cancer which are at various stages of development, may give insight into their changing chemical compositions and therefore enable the interpretation of internal mechanisms. Another useful study would be look at various types of cancer and contrast the wavenumbers which are found to be key for discrimination, to ascertain is the discriminators vary from cancer to cancer or if they are consistent.

A second interesting possibility would be to combine the MA code with a form of cluster analysis (CA). As MA is a supervised learning algorithm it has to be supplied with predefined spectra to characterise and learn from. It was demonstrated in the Chapter 4 that the quality of the training set is key and that it is possible for tissue samples to be miss labelled. CA is an unsupervised ML algorithm which groups the spectra in such a way that the spectra within a cluster are more similar compared to the spectra outside the cluster. One possibility would be to use CA to label the tissue images used for the learning process rather than a skilled histopathologist. This should result in the spectra used within each of the training datasets being similar in structure. The entire CA-MA algorithm would therefore be automated with no 'human' intervention at any stage besides collecting the data. Care would have to be taken after the CA-MA algorithm had run to try and understand which of the spectra had been grouped in the CA as they may not necessarily be the desired sample types. This type of algorithm may be ideal if trying to study a very complicated tissue sample which has poorly defined tissue boundaries, which would be a problem for labelling by a person.

The current progression of SNOM instruments in general is in adopting the newly developed quantum cascade lasers (QCL). Aperture SNOMs used in conjunction with a IR-QCL source have not yet been demonstrated in any experimental studies but they appear to be a very promising alternative to the costly and complicated IR-FELs. Their beam characteristics are very different to the FEL as the peak power is considerably lower but as the repetition rate is much higher it is capable of achieving similar average powers. The QCL signal measured by the MCT detector would be very small, so the SNOM would need to accommodate a lock-in amplifier which would extract the weaker signal from the noise. A QCL-SNOM instrument would benefit from an improved imaging speed as the SNOM is no longer locked to the FELs repetition rate of 10 Hz.

An added advantage of using a QCL is that it is very efficient at sweeping through multiple wavenumbers, being able to sweep through the 1000-1800 cm^{-1} range many times a second with a very high spectral resolution. This would allow the SNOM to collect a full spectrum at each pixel position rather than just a single wavelength as it does currently. The output of such an instrument would be akin to a FTIR datacube as it contains both the spatial variation and spectral variation of the sample. This makes the SNOM a much more versatile instrument as it can be used to collect many wavenumbers over a longer scan period ($\approx$ 1 hour) or scan a few wavenumbers very quickly ($\approx$ minutes). All of this while still retaining the high spatial resolution of the SNOM instrument. The QCL allows for a desktop instrument which makes the SNOM much more convenient compared to only being able to use it at FEL sites.

If the SNOM were to create an output akin to a datacube it would allow for ML techniques such as MA to study the spectra taken at much higher spatial resolutions. This would allow for not just sample discrimination but potentially the labelling of structures within the cell. If this were possible the imaging of organelles and molecular distributions within cells at different diseased stages would be very interesting.

Bibliography

[1] Christopher J.L. Lozano, Murray. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859):2095–2128, 2012. ISSN 1474547X. doi: 10.1016/S0140-6736(12)61728-0.

[2] cancer rate increase. URL `http://www.who.int/mediacentre/news/releases/2003/pr27/en/`.

[3] Michael Roerecke, Kevin D Shield, Susumu Higuchi, Atsushi Yoshimura, and Elisabeth Larsen. Systematic reviews Estimates of alcohol-related oesophageal cancer burden in Japan : systematic review and meta-analyses. (May 2014):329–338, 2015.

[4] Cancer research UK. How we spend your money. URL `http://www.cancerresearchuk.org/how-we-spend-your-money`.

[5] Edwards BK Horner MJ, Ries LAG, Krapcho M, Neyman N, Aminou R, Howlader N, Altekruse SF, Feuer EJ, Huang L, Mariotto A, Miller BA, Lewis DR, Eisner MP, Stinchcomb DG. SEER Cancer Statistics Review, 1975-2006. *National Cancer Institute*, 2009. URL `https://seer.cancer.gov/csr/1975{_}2006/`.

[6] Cancer Research UK. Breast cancer survival statistics. URL `http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/survival{#}By`.

[7] a Sudhakar. History of Cancer, Ancient and Modern Treatment Methods. *J Cancer Sci Ther.*, 1(2):1–4, 2010. ISSN 1948-5956. doi: 10.4172/1948-5956.

100000e2.History. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927383/pdf/nihms226784.pdf.

[8] Gabriele Bergers, Rolf Brekken, Gerald McMahon, H Thiennu, Itoh Takeshi, Tamaki Kazuhiko, Kazuhiko Tanzawa, Philip Thorpe, Shigeyoshi Itohara, Zena Werb, and Douglas Hanahan. Matrix Metalloproteinase-9 Triggres the Angiogenic Switch During Carcinogenesis. 2(October):20–25, 2000.

[9] Preetha Anand, Ajaikumar B. Kunnumakara, Chitra Sundaram, Kuzhuvelil B. Harikumar, Sheeja T. Tharakan, Oiki S. Lai, Bokyung Sung, and Bharat B. Aggarwal. Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical Research*, 25(9):2097–2116, 2008. ISSN 07248741. doi: 10.1007/s11095-008-9661-9.

[10] Ahmedin Jemal, Freddie Bray, and Jacques Ferlay. Global Cancer Statistics: 2011. *CA Cancer J Clin*, 49(2):1,33–64, 1999. ISSN 0007-9235. doi: 10.3322/caac.20107.Available. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=10200776.

[11] Michael Pilling and Peter Gardner. Fundamental developments in infrared spectroscopic imaging for biomedical applications. *Chem. Soc. Rev.*, 45(7):1935–1957, 2016. ISSN 0306-0012. doi: 10.1039/C5CS00846H. URL http://xlink.rsc.org/?DOI=C5CS00846H.

[12] Lord Carter of Coles. Report of the review of NHS pathlogy services in England. pages 1–51, 2006. URL http://webarchive.nationalarchives.gov.uk/20130107105354/http:/www.dh.gov.uk/prod{_}consum{_}dh/groups/dh{_}digitalassets/@dh/@en/documents/digitalasset/dh{_}091984.pdf.

[13] Thomas A. Stamey. Biological Determinants of Cancer Progression in Men With Prostate Cancer. *Jama*, 281(15):1395, 1999. ISSN 0098-7484. doi: 10.1001/jama.281.15.1395. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.281.15.1395.

[14] Marlena Schoenberg Fejzo and Dennis J Slamon. *Tissue Microarrays from Frozen Tissues-OCT Technique*, pages 73–80. Humana Press, Totowa, NJ, 2010. ISBN 978-1-60761-806-5. doi: 10.1007/978-1-60761-806-5_8. URL https://doi.org/10.1007/978-1-60761-806-5{_}8.

[15] NHS. Biopsy - NHS. URL https://www.nhs.uk/conditions/biopsy/.

[16] R D Odze. Diagnosis and grading of dysplasia in Barrett's oesophagus. *Journal of Clinical Pathology*, 59(10):1029–1038, 2006. ISSN 0021-9746. doi: 10.1136/jcp.2005.035337. URL http://jcp.bmj.com/cgi/doi/10.1136/jcp.2005.035337.

[17] J. B. Lattouf and F. Saad. Gleason score on biopsy: Is it reliable for predicting the final grade on pathology? *BJU International*, 90(7):694–698, 2002. ISSN 14644096. doi: 10.1046/j.1464-410X.2002.02990.x.

[18] Daniel C. Paech, Adèle R. Weston, Nick Pavlakis, Anthony Gill, Narayan Rajan, Helen Barraclough, Bronwyn Fitzgerald, and Maximiliano Van Kooten. A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer. *Journal of Thoracic Oncology*, 6(1):55–63, 2011. ISSN 15561380. doi: 10.1097/JTO.0b013e3181fc0878.

[19] Jin Tae Kwak, Stephen M. Hewitt, Saurabh Sinha, and Rohit Bhargava. Multimodal microscopy for automated histologic analysis of prostate cancer. *BMC Cancer*, 11(1):62, 2011. ISSN 14712407. doi: 10.1186/1471-2407-11-62. URL http://www.biomedcentral.com/1471-2407/11/62.

[20] Nicole J. Kline and Patrick J. Treado. Raman Chemical Imaging of Breast Tissue. *Journal of Raman Spectroscopy*, 28(2-3):119–124, 1997. ISSN 0377-0486. doi: 10.1002/(SICI)1097-4555(199702)28:2/3⟨119::AID-JRS73⟩3.0.CO;2-3. URL http://doi.wiley.com/10.1002/{%}28SICI{%}291097-4555{%}28199702{%}2928{%}3A2/3{%}3C119{%}3A{%}3AAID-JRS73{%}3E3.0.CO{%}3B2-3.

[21] Norbert Bergner, Bernd F. M. Romeike, Rupert Reichart, Rolf Kalff, Christoph Krafft, and Jürgen Popp. Tumor margin identification and prediction of the primary tumor from brain metastases using FTIR imaging and support vector machines. *The Analyst*, 138(14):3983, 2013. ISSN 0003-2654. doi: 10.1039/c3an00326d. URL `http://xlink.rsc.org/?DOI=c3an00326d`.

[22] Laven Mavarani, Dennis Petersen, Samir F. El-Mashtoly, Axel Mosig, Andrea Tannapfel, Carsten Kötting, and Klaus Gerwert. Spectral histopathology of colon cancer tissue sections by Raman imaging with 532 nm excitation provides label free annotation of lymphocytes, erythrocytes and proliferating nuclei of cancer cells. *The Analyst*, 138(14):4035, 2013. ISSN 0003-2654. doi: 10.1039/c3an00370a. URL `http://xlink.rsc.org/?DOI=c3an00370a`.

[23] Paul Bassan, Joe Mellor, Jonathan Shapiro, Kaye J Williams, Michael P Lisanti, and Peter Gardner. Transmission FT-IR Chemical Imaging on Glass Substrates: Applications in Infrared Spectral Histopathology. 2014. doi: 10.1021/ac403412n.

[24] Michael J Pilling, Alex Henderson, Jonathan H Shanks, Michael D Brown, Noel W Clarke, and Peter Gardner. Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation. *The Analyst*, 142(8):1258–1268, 2017. ISSN 0003-2654. doi: 10.1039/c6an02224c. URL `http://www.ncbi.nlm.nih.gov/pubmed/27921102`.

[25] Ehsan Gazi, John Dwyer, Nicholas Lockyer, Peter Gardner, John C. Vickerman, Jaleel Miyan, Claire A. Hart, Mick Brown, Jonathan H. Shanks, and Noel Clarke. The combined application of FTIR microspectroscopy and ToF-SIMS imaging in the study of prostate cancer. *Faraday Discussions*, 126:41, 2004. ISSN 1359-6640. doi: 10.1039/b304883g. URL `http://xlink.rsc.org/?DOI=b304883g`.

[26] Charles H Camp Jr, Young Jong Lee, John M Heddleston, Christopher M Hartshorn, Angela R Hight Walker, Jeremy N Rich, Justin D Lathia, and

Marcus T Cicerone. High-speed coherent Raman Fingerpring Imaging of Biological Tissues. *Nature Photonics*, 8:627–634, 2014. ISSN 0036-8075. doi: 10.1038/nphoton.2014.145.High-Speed.

[27] Rohit Bhargava. Infrared spectroscopic imaging: The next generation. *Applied Spectroscopy*, 66(10):1091–1120, 2012. ISSN 00037028. doi: 10.1366/12-06801.

[28] J. Depciuch, E. Kaznowska, S. Golowski, A. Koziorowska, I. Zawlik, M. Cholewa, K. Szmuc, and J. Cebulski. Monitoring breast cancer treatment using a Fourier transform infrared spectroscopy-based computational model. *Journal of Pharmaceutical and Biomedical Analysis*, 143:261–268, 2017. ISSN 1873264X. doi: 10.1016/j.jpba.2017.04.039. URL `http://dx.doi.org/10.1016/j.jpba.2017.04.039`.

[29] Wenyue Yang, Xilin Xiao, Jun Tan, and Qingyun Cai. In situ evaluation of breast cancer cell growth with 3D ATR-FTIR spectroscopy. *Vibrational Spectroscopy*, 49(1):64–67, 2009. ISSN 09242031. doi: 10.1016/j.vibspec.2008.04.016.

[30] Mark A. Mackanos and Christopher H. Contag. FTIR microspectroscopy for improved prostate cancer diagnosis. *Trends in Biotechnology*, 27(12):661–663, 2009. ISSN 01677799. doi: 10.1016/j.tibtech.2009.09.001.

[31] H. Ukkonen, S. Kumar, J. Mikkonen, T. Salo, S. P. Singh, A. P. Koistinen, E. Goormaghtigh, and A. M. Kullaa. Changes in the microenvironment of invading melanoma and carcinoma cells identified by FTIR imaging. *Vibrational Spectroscopy*, 79:24–30, 2015. ISSN 09242031. doi: 10.1016/j.vibspec.2015.04.005. URL `http://dx.doi.org/10.1016/j.vibspec.2015.04.005`.

[32] G I Dovbeshko, N Y Gridina, E B Kruglova, and O P Pashchuk. FTIR spectroscopy studies of nucleic acid damage. *Talanta*, 53:233–246, 2000. ISSN 1873-3573.

[33] Jayakrupakar Nallala, Gavin Rhys Lloyd, Michael Hermes, Neil Shepherd, and Nick Stone. Enhanced spectral histology in the colon using

high-magnification benchtop FTIR imaging. *Vibrational Spectroscopy*, 91: 83–91, 2017. ISSN 09242031. doi: 10.1016/j.vibspec.2016.08.013. URL http://dx.doi.org/10.1016/j.vibspec.2016.08.013.

[34] Shady G El-Tawil, Rohana Adnan, Zaki N Muhamed, and Nor Hayati Othman. Comparative study between Pap smear cytology and FTIR spectroscopy: a new tool for screening for cervical cancer. *Pathology*, 40(6): 600–3, 2008. ISSN 0031-3025. doi: 10.1080/00313020802320622. URL http://www.ncbi.nlm.nih.gov/pubmed/18752127.

[35] B. R. Wood, L. Chiriboga, H. Yee, M. A. Quinn, D. McNaughton, and M. Diem. Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. *Gynecologic Oncology*, 93(1):59–68, 2004. ISSN 00908258. doi: 10.1016/j.ygyno.2003.12.028.

[36] S. Mordechai, R. K. Sahu, Z. Hammody, S. Mark, K. Kantarovich, H. Guterman, A. Podshyvalov, J. Goldstein, and S. Argov. Possible common biomarkers from FTIR microspectroscopy of cervical cancer and melanoma. *Journal of Microscopy*, 215(1):86–91, 2004. ISSN 00222720. doi: 10.1111/j.0022-2720.2004.01356.x.

[37] Ehsan Gazi, Matthew Baker, John Dwyer, Nicholas P. Lockyer, Peter Gardner, Jonathan H. Shanks, Roy S. Reeve, Claire A. Hart, Noel W. Clarke, and Michael D. Brown. A Correlation of FTIR Spectra Derived from Prostate Cancer Biopsies with Gleason Grade and Tumour Stage. *European Urology*, 50(4):750–761, 2006. ISSN 03022838. doi: 10.1016/j.eururo.2006.03.031.

[38] Paul Bassan, Ashwin Sachdeva, Jonathan H. Shanks, Mick D. Brown, Noel W. Clarke, and Peter Gardner. Whole organ cross-section chemical imaging using label-free mega-mosaic FTIR microscopy. *The Analyst*, 138(23):7066, 2013. ISSN 0003-2654. doi: 10.1039/c3an01674a. URL http://xlink.rsc.org/?DOI=c3an01674a.

[39] Shashi D. Buluswar and Bruce A. Draper. Color machine vision for autonomous vehicles. *Engineering Applications of Artificial Intelligence*, 11 (2):245–256, 1998. ISSN 09521976. doi: 10.1016/S0952-1976(97)00079-1.

[40] Yifei Wang, Yuan Gao, Alin Achim, and Naim Dahnoun. Robust obstacle detection based on a novel disparity calculation method and G-disparity. *Computer Vision and Image Understanding*, 123:23–40, 2014. ISSN 1090235X. doi: 10.1016/j.cviu.2014.02.014. URL `http://dx.doi.org/10.1016/j.cviu.2014.02.014`.

[41] Lijuan Cao and Francis E. H. Tay. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transaction on Neural Networks*, 14(6):1506–18, 2003. ISSN 1045-9227. doi: 10.1109/TNN.2003.820556.

[42] Wei Huang, Yoshiteru Nakamori, and Shou Yang Wang. Forecasting stock market movement direction with support vector machine. *Computers and Operations Research*, 32(10):2513–2522, 2005. ISSN 03050548. doi: 10.1016/j.cor.2004.03.016.

[43] Constantine Papageorgiou and Tomaso Poggio. Trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000. ISSN 09205691. doi: 10.1023/A:1008162616689.

[44] Claus Bahlmann, Amar Patel, Jeffrey Johnson, Jie Ni, Andrei Chekkoury, Parmeshwar Khurd, Ali Kamen, Leo Grady, Elizabeth Krupinski, Anna Graham, and Ronald Weinstein. Automated detection of diagnostically relevant regions in H&E stained digital pathology slides. *Proc. of SPIE, medical imaging*, 831504(February 2012):831504, 2012. ISSN 16057422. doi: 10.1117/12.912484.

[45] Lena Gorelick, Olga Veksler, Mena Gaed, Jose A. Gomez, Madeleine Moussa, Glenn Bauman, Aaron Fenster, and Aaron D. Ward. Prostate histopathology: Learning tissue component histograms for cancer detection and classification. *IEEE Transactions on Medical Imaging*, 32(10):1804–1818, 2013. ISSN 02780062. doi: 10.1109/TMI.2013.2265334.

[46] Benjamin R. Smith, Katherine M. Ashton, Andrew Brodbelt, Timothy Dawson, Michael D. Jenkinson, Neil T. Hunt, David S. Palmer, and Matthew J. Baker. Combing Random Forest and 2D Correlations Analysis to Identify Serum Spectral Signatured for Neuroncology. *Analyst*, pages 1–8, 2016. doi: 10.1039/b000000x/2.

[47] M J Baker, E Gazi, M D Brown, J H Shanks, P Gardner, and N W Clarke. FTIR-based spectroscopic analysis in the identification of clinically aggressive prostate cancer. *British Journal of Cancer*, 99(11):1859–1866, 2008. ISSN 0007-0920. doi: 10.1038/sj.bjc.6604753. URL `http://www.nature.com/doifinder/10.1038/sj.bjc.6604753`.

[48] Ketan Gajjar, Júlio Trevisan, Gemma Owens, Patrick J. Keating, Nicholas J. Wood, Helen F. Stringfellow, Pierre L. Martin-Hirsch, and Francis L. Martin. Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. *The Analyst*, 138(14):3917, 2013. ISSN 0003-2654. doi: 10.1039/c3an36654e. URL `http://xlink.rsc.org/?DOI=c3an36654e`.

[49] Peter Lasch, Wolfgang Haensch, Dieter Naumann, and Max Diem. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1688 (2):176–186, 2004. ISSN 09254439. doi: 10.1016/j.bbadis.2003.12.006.

[50] Izabela Zawlik, Ewa Kaznowska, Jozef Cebulski, Magdalena Kolodziej, Joanna Depciuch, Jitraporn Vongsvivut, and Marian Cholewa. FPA-FTIR Microspectroscopy for Monitoring Chemotherapy Efficacy in Triple-Negative Breast Cancer. *Scientific Reports*, 6:1–8, 2016. ISSN 20452322. doi: 10.1038/srep37333. URL `http://dx.doi.org/10.1038/srep37333`.

[51] E. Kaznowska, J. Depciuch, K. Szmuc, and J. Cebulski. Use of FTIR spectroscopy and PCA-LDC analysis to identify cancerous lesions within the human colon. *Journal of Pharmaceutical and Biomedical Analysis*, 134: 259–268, 2017. ISSN 1873264X. doi: 10.1016/j.jpba.2016.11.047. URL http://dx.doi.org/10.1016/j.jpba.2016.11.047.

[52] Heinz Fabian, Ngoc Anh Ngo Thi, Michael Eiden, Peter Lasch, Jürgen Schmitt, and Dieter Naumann. Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1758(7):874–882, 2006. ISSN 00052736. doi: 10.1016/j.bbamem.2006.05.015. URL http://linkinghub.elsevier.com/retrieve/pii/S0005273606001994.

[53] Matthew J. Baker, Hugh J. Byrne, John Chalmers, Peter Gardner, Royston Goodacre, Alex Henderson, Sergei G. Kazarian, Francis L. Martin, Julian Moger, Nick Stone, and Josep Sulé-Suso. Clinical applications of infrared and Raman spectroscopy: state of play and future challenges. *The Analyst*, (1985):1735–1757, 2018. ISSN 0003-2654. doi: 10.1039/C7AN01871A. URL http://xlink.rsc.org/?DOI=C7AN01871A.

[54] Hugh J. Byrne, Malgorzata Baranska, Gerwin J. Puppels, Nick Stone, Bayden Wood, Kathleen M. Gough, Peter Lasch, Phil Heraud, Josep Sulé-Suso, and Ganesh D. Sockalingum. Spectropathology for the next generation: Quo vadis? *The Analyst*, 140(7):2066–2073, 2015. ISSN 0003-2654. doi: 10.1039/C4AN02036G. URL http://xlink.rsc.org/?DOI=C4AN02036G.

[55] H Heinzelmann and D W Pohl. Scanning near-field optical microscopy. *Applied Physics A*, 59(2):89–101, 1994. ISSN 1432-0630. doi: 10.1007/BF00332200. URL http://dx.doi.org/10.1007/BF00332200.

[56] D. W. Pohl, W. Denk, and M. Lanz. Optical stethoscopy: Image recording with resolution $\lambda/20$. *Applied Physics Letters*, 44(7):651–653, 1984. ISSN 00036951. doi: 10.1063/1.94865.

[57] Marc Richter, Martin Hedegaard, Tanja Deckert-Gaudig, Peter Lampen, and Volker Deckert. Laterally resolved and direct spectroscopic evidence of nanometer-sized lipid and protein domains on a single cell. *Small*, 7(2): 209–214, 2011. ISSN 16136810. doi: 10.1002/smll.201001503.

[58] Bayden R. Wood, Elena Bailo, Mehdi Asghari Khiavi, Leann Tilley, Samantha Deed, Tanja Deckert-Gaudig, Don McNaughton, and Volker Deckert. Tip-enhanced raman scattering (TERS) from hemozoin crystals within a sectioned erythrocyte. *Nano Letters*, 11(5):1868–1873, 2011. ISSN 15306984. doi: 10.1021/nl103004n.

[59] A. Cricenti, R. Generosi, M. Luce, P. Perfetti, G. Margaritondo, D. Talley, J. S. Sanghera, I. D. Aggarwal, and N. H. Tolk. Very high resolution near-field chemical imaging using an infrared free electron laserPresented at the LANMAT 2001 Conference on the Interaction of Laser Radiation with Matter at Nanoscopic Scales: From Single Molecule Spectroscopy to Materials Processing, Venice, 36 October, 2001. *Physical Chemistry Chemical Physics*, 4(12):2738–2741, 2002. ISSN 14639076. doi: 10.1039/b109279k. URL `http://xlink.rsc.org/?DOI=b109279k`.

[60] Diane E. Halliwell, Camilo L M Morais, Kássio M G Lima, Julio Trevisan, Michele R F Siggel-King, Tim Craig, James Ingham, David S. Martin, Kelly A. Heys, Maria Kyrgiou, Anita Mitra, Evangelos Paraskevaidis, Georgios Theophilou, Pierre L. Martin-Hirsch, Antonio Cricenti, Marco Luce, Peter Weightman, and Francis L. Martin. Imaging cervical cytology with scanning near-field optical microscopy (SNOM) coupled with an IR-FEL. *Scientific Reports*, 6(June):1–11, 2016. ISSN 20452322. doi: 10.1038/srep29494.

[61] P Perner, A Rapp, C Dressler, L Wollweber, J Beuthan, K O Greulich, and M Hausmann. Variations in cell surfaces of oestrogen treated breast cancer cells detected by a combined instrument for far-field and near-field microscopy. *Analytical Cellular Pathology*, 24:89–100, 2002.

[62] Gerd Kaupp. Scanning near-field optical microscopy on rough surfaces: Applications in chemistry, biology, and medicine. *International Journal of Photoenergy*, 2006:1–22, 2006. ISSN 1110662X. doi: 10.1155/IJP/2006/69878.

[63] D Forman. Review article: oesophago-gastric adenocarcinoma – an epidemiological perspective. *Alimentary pharmacology & therapeutics*, 20 Suppl 5: 55–60; discussion 61–2, 2004. ISSN 0269-2813. doi: 10.1111/j.1365-2036.2004.02133.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/15456465`.

[64] Elizabeth Montgomery, John R. Goldblum, Joel K. Greenson, Marian M. Haber, Laura W. Lamps, Gregory Y. Lauwers, Audrey J. Lazenby, David N. Lewin, Marie E. Robert, Kay Washington, Marianna L. Zahurak, and John Hart. Dysplasia as a predictive marker for invasive carcinoma in barrett esophagus: A follow-up study based on 138 cases from a diagnostic variability study. *Human Pathology*, 32(4):379–388, 2001. ISSN 00468177. doi: 10.1053/hupa.2001.23511.

[65] Z. Hammody, S. Argov, R. K. Sahu, E. Cagnano, R. Moreh, and S. Mordechai. Distinction of malignant melanoma and epidermis using IR micro-spectroscopy and statistical methods. *The Analyst*, 133(3):372, 2008. ISSN 0003-2654. doi: 10.1039/b712040k. URL `http://xlink.rsc.org/?DOI=b712040k`.

[66] Stuart A.C. McDonald, Trevor A. Graham, Danielle L. Lavery, Nicholas A. Wright, and Marnix Jansen. The Barrett's Gland in Phenotype Space. *Cmgh*, 1(1):41–54, 2015. ISSN 2352345X. doi: 10.1016/j.jcmgh.2014.10.001. URL `http://dx.doi.org/10.1016/j.jcmgh.2014.10.001`.

[67] N. J. Shaheen, M. A. Crosby, E. M. Bozymski, and R. S. Sandler. Is there publication bias in the reporting of cancer risk in Barrett's esophagus? *Gastroenterology*, 119(2):333–338, 2000. ISSN 00165085. doi: 10.1053/gast.2000.9302.

[68] a. D. Smith, M. R F Siggel-King, G. M. Holder, a. Cricenti, M. Luce, P. Harrison, D. S. Martin, M. Surman, T. Craig, S. D. Barrett, a. Wolski, D. J.

Dunning, N. R. Thompson, Y. Saveliev, D. M. Pritchard, a. Varro, S. Chattopadhyay, and P. Weightman. Near-field optical microscopy with an infrared free electron laser applied to cancer diagnosis. *Applied Physics Letters*, 102(5):1–5, 2013. ISSN 00036951. doi: 10.1063/1.4790436.

[69] Timothy Craig, Andrew D. Smith, Gareth M. Holder, James Ingham, Caroline I. Smith, Andrea Varro, D. Mark Pritchard, Steve D. Barrett, David S. Martin, Paul Harrison, Andrzej Wolski, Antonio Cricenti, Marco Luce, Mark Surman, Swapan Chattopadhyay, Peter Weightman, and Michele R.F. Siggel-King. SNOM Imaging of a Crypt-Like Feature in Adenocarcinoma Associated with Barrett's Oesophagus. *Physica Status Solidi (B) Basic Research*, 1700518:1–7, 2018. ISSN 15213951. doi: 10.1002/pssb.201700518.

[70] Luca Quaroni and Alan G. Casson. Characterization of Barrett esophagus and esophageal adenocarcinoma by Fourier-transform infrared microscopy. *The Analyst*, 134(6):1240, 2009. ISSN 0003-2654. doi: 10.1039/b823071d. URL `http://xlink.rsc.org/?DOI=b823071d`.

[71] Jian-Sheng Wang, Jing-Sen Shi, Yi-Zhuang Xu, Xiao-Yi Duan, Li Zhang, Jing Wang, Li-Ming Yang, Shi-Fu Weng, and Jin-Guang Wu. FT-IR spectroscopic analysis of normal and cancerous tissues of esophagus. *World J Gastroenterol*, 9(9):1897–9, 2003. ISSN 1007-9327. doi: 10.3748/WJG.V9.I9.1897. URL `http://www.ncbi.nlm.nih.gov/pubmed/12970871`.

[72] Donna E. Maziak, Minh T. Do, Farid M. Shamji, Sudhir R. Sundaresan, D. Garth Perkins, and Patrick T.T. Wong. Fourier-transform infrared spectroscopic study of characteristic molecular structure in cancer cells of esophagus: An exploratory study. *Cancer Detection and Prevention*, 31(3):244–253, 2007. ISSN 0361090X. doi: 10.1016/j.cdp.2007.03.003.

[73] Oliver Old, Gavin Lloyd, Martin Isabelle, L Max Almond, Catherine Kendall, Karol Baxter, Neil Shepherd, Angela Shore, Nick Stone, and Hugh Barr. Automated cytological detection of Barrett ' s neoplasia with infrared spectroscopy. *Journal of Gastroenterology*, 53(2):227–235, 2018. ISSN 1435-5922. doi: 10.1007/s00535-017-1344-z.

[74] Timothy Craig. *The development of infrared scanning near-field optical microscopy for the study of cancer and other biological problems*. PhD thesis, University of Liverpool, 2016.

[75] E Abbe. Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung. *M. Schultze's Archiv für mikroskopische Anatomie*, 9:413–468, 1873.

[76] Stefanie E. Glassford, Bernadette Byrne, and Sergei G. Kazarian. Recent applications of ATR FTIR spectroscopy and imaging to proteins. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1834(12):2849–2858, 2013. ISSN 15709639. doi: 10.1016/j.bbapap.2013.07.015. URL `http://dx.doi.org/10.1016/j.bbapap.2013.07.015`.

[77] J. Madejová. FTIR techniques in clay mineral studies. *Vibrational Spectroscopy*, 31(1):1–10, 2003. ISSN 09242031. doi: 10.1016/S0924-2031(02)00065-6.

[78] Quantification of minerals from ATR-FTIR spectra with spectral interferences using the MRC method. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 181:7–12, 2017. ISSN 13861425. doi: 10.1016/j.saa.2017.02.012. URL `http://dx.doi.org/10.1016/j.saa.2017.02.012`.

[79] P Luc and S Gerstenkorn. Fourier transform spectroscopy in the visible and ultraviolet range. *Appl. Opt.*, 17(9):1327–1331, may 1978. doi: 10.1364/AO.17.001327. URL `http://ao.osa.org/abstract.cfm?URI=ao-17-9-1327`.

[80] Robert K Chan, P K Lim, Xuzhu Wang, and M H Chan. Fourier transform ultraviolet-visible spectrometer based on a beam-folding technique. *Opt. Lett.*, 31(7):903–905, apr 2006. doi: 10.1364/OL.31.000903. URL `http://ol.osa.org/abstract.cfm?URI=ol-31-7-903`.

[81] N. De Oliveira, D. Joyeux, D. Phalippou, J. C. Rodier, F. Polack, M. Vervloet, and L. Nahon. A Fourier transform spectrometer without a beam splitter for the vacuum ultraviolet range: From the optical design to

the first UV spectrum. *Review of Scientific Instruments*, 80(4), 2009. ISSN 00346748. doi: 10.1063/1.3111452.

[82] Abiodun Ogunleke, Vladimir Bobroff, Hsiang-Hsin Chen, Jeremy Rowlette, Maylis Delugin, Benoit Recur, Yeukuang Hwu, and Cyril Petibois. Fourier-transform vs. quantum-cascade-laser infrared microscopes for histopathology: From lab to hospital? *TrAC Trends in Analytical Chemistry*, 89:190–196, 2017. ISSN 01659936. doi: 10.1016/j.trac.2017.02.007. URL `http://linkinghub.elsevier.com/retrieve/pii/S0165993616303387`.

[83] Benjamin Bird and Matthew J. Baker. Quantum Cascade Lasers in Biomedical Infrared Imaging. *Trends in Biotechnology*, 33(10):557–558, 2015. ISSN 18793096. doi: 10.1016/j.tibtech.2015.07.003. URL `http://dx.doi.org/10.1016/j.tibtech.2015.07.003`.

[84] P. B. Fellgett. The nature and origin of multiplex fourier spectrometry. *Notes and Records of the Royal Society of London*, 60(1):91–93, 2006. ISSN 00359149. URL `http://www.jstor.org/stable/20462556`.

[85] N Cortadellas I Raméntol. Transmission electron microscopy in cell biology: sample preparation techniques and image information. In *Handbook of instrumental techniques from CCiTUB*. 2012. URL `http://diposit.ub.edu/dspace/handle/2445/31942`.

[86] E.H.Synge. A suggested method for extending microscopic resolution into the ultra-microscopic region. *PHILOSOPHICAL MAGAZINE*, 6(35):356–362, 1928.

[87] E. Betzig, A. Lewis, A. Harootunian, M. Isaacson, and E. Kratschmer. Near Field Scanning Optical Microscopy (NSOM). *Biophysical Journal*, 49(1): 269–279, 1986. ISSN 00063495. doi: 10.1016/S0006-3495(86)83640-2.

[88] G. Nicholls E.A.Ash. Super-resolution Aperture Scanning Microscope. *Nature*, 237(5357):510–512, 1972.

[89] H. Rohrer G. Binnig. Surface Studies by Scanning Tunneling Microscopy. *Physics Review Letterss*, 49(1):57–61, 1982.

[90] J. Stadler B. Yeo. Tip-enhanced Raman Spectroscopy Its status, challenges and future directions. *Chemical Physics Letters*, 472(2):1–13, 2009.

[91] Raoul M Stöckle, Yung Doug Suh, Volker Deckert, and Renato Zenobi. Nanoscale chemical analysis by tip-enhanced Raman spectroscopy. *Chemical Physics Letters*, 318(1):131–136, 2000. ISSN 0009-2614. doi: http://dx.doi.org/10.1016/S0009-2614(99)01451-7. URL `http://www.sciencedirect.com/science/article/pii/S0009261499014517`.

[92] S.A.Asselborn Y.V. Miklyaev. Superresolution microscopy in far-field by near-field optical random mapping nanoscopy. *APPLIED PHYSICS LETTERS*, 105(11), 2014. doi: 10.1063/1.4895922.

[93] HAROOTUNIAN. A. SUPERRESOLUTION FLUORESCENCE NEAR-FIELD SCANNING OPTICAL MICROSCOPY. *PPLIED PHYSICS LETTERS*, 49(11):674–676, 1986.

[94] Lothar Schermelleh, Rainer Heintzmann, and Heinrich Leonhardt. A guide to super-resolution fluorescence microscopy. *Journal of Cell Biology*, 190(2): 165–175, 2010. ISSN 00219525. doi: 10.1083/jcb.201002018.

[95] Bo Huang, Hazen Babcock, and Xiaowei Zhuang. Breaking the diffraction barrier: Super-resolution imaging of cells. *Cell*, 143(7):1047–1058, 2010. ISSN 00928674. doi: 10.1016/j.cell.2010.12.002. URL `http://dx.doi.org/10.1016/j.cell.2010.12.002`.

[96] Eric Betzig and Robert J. Chichester. Single Molecules Observed by Near-Field Scanning Optical Microscopy. *Science*, 262(5138), 1993.

[97] D. Vobornik, G. Margaritondo, J. S. Sanghera, P. Thielen, I. D. Aggarwal, B. Ivanov, N. H. Tolk, V. Manni, S. Grimaldi, A. Lisi, S. Rieti, D. W. Piston, R. Generosi, M. Luce, P. Perfetti, and A. Cricenti. Spectroscopic infrared scanning near-field optical microscopy (IR-SNOM). *Journal of Alloys and*

Compounds, 401(1-2):80–85, 2005. ISSN 09258388. doi: 10.1016/j.jallcom.2005.02.057.

[98] N. R. Thompson, D. J. Dunning, J. A. Clarke, M. Surman, A. D. Smith, Y. Saveliev, and S. Leonard. First lasing of the ALICE infra-red Free-Electron Laser. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 680:117–123, 2012. ISSN 01689002. doi: 10.1016/j.nima.2012.02.049. URL http://dx.doi.org/10.1016/j.nima.2012.02.049.

[99] D. A. G. Deacon, J. M. J. Madey, and L. R. Elias. First Operation of a Free-Electron Laser. (16):16–18, 1977.

[100] Herman Winick and Stanford Linear. Fourth generation light sources. *Conf.Proc.*, C970512:37–41, 1997.

[101] L B Jones, J W McKenzie, K J Middleman, B L Militsyn, Y M Saveliev, and S L Smith. The ALICE Energy Recovery Linac Project overview and injector performance. *Journal of Physics: Conference Series*, 298:012007, 2011. ISSN 1742-6596. doi: 10.1088/1742-6596/298/1/012007. URL http://stacks.iop.org/1742-6596/298/i=1/a=012007?key=crossref.29310e41cea765f4a15ba2d21462056d.

[102] *ALICE Safety Handbook*. 2011. URL http://projects.astec.ac.uk/ERLPManual/images/c/c4/ALICE{_}Safety{_}Handbook.pdf.

[103] W. B. Colson. Theory of a free electron laser. *Physics Letters*, 59A(3):187–190, 1976. ISSN 03759601. doi: 10.1016/0375-9601(76)90561-2. URL http://www.sciencedirect.com/science/article/pii/0375960176905612.

[104] Olympus. IX3 Inverted microscope. URL https://www.olympus-lifescience.com/en/microscopes/inverted/ix3-icsi/.

[105] D. B. Talley, L. B. Shaw, J. S. Sanghera, I. D. Aggarwal, A. Cricenti, R. Generosi, M. Luce, G. Margaritondo, J. M. Gilligan, and N. H. Tolk. Scanning

near field infrared microscopy using chalcogenide fiber tips. *Materials Letters*, 42(5):339–344, 2000. ISSN 0167577X. doi: 10.1016/S0167-577X(99)00201-3.

[106] CorActive. IR fibres. URL `http://coractive.com/products/mid-ir-fibers-lasers/select-cutoff-singlemode-fiber/index.html`.

[107] M A Unger, D A Kossakovski, R Kongovi, J L Beauchamp, J D Baldeschwieler, and D V Palanker. Etched chalcogenide fibers for near-field infrared scanning microscopy. *Review of Scientific Instruments*, 69(8):2988–2993, 1998. ISSN 00346748. doi: 10.1063/1.1149045. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-0032133552{&}partnerID=40{&}md5=6ca899273206775f67c551dc95ec2f31`.

[108] Chia Yun Chang, Ming Tsung Hsu, Emilio Xavier Esposito, and Yufeng J. Tseng. Oversampling to overcome overfitting: Exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods. *Journal of Chemical Information and Modeling*, 53(4):958–971, 2013. ISSN 15499596. doi: 10.1021/ci4000536.

[109] R Bruch, N Afanasyeva, and S Gummuluri. Analysis and classification of normal and pathological skin tissue spectra using neural networks. *Subsurface Sensing Technologies and Applications Ii*, 4129(July 2000):196–206, 2000. doi: 10.1117/12.390617.

[110] Cungui Cheng, Wei Xiong, and Yumei Tian. Classification of rat FTIR colon cancer data using wavelets and BPNN. *Chinese Journal of Chemistry*, 27 (5):911–914, 2009. ISSN 1001604X. doi: 10.1002/cjoc.200990154.

[111] Paul Bassan, Hugh J. Byrne, Franck Bonnier, Joe Lee, Paul Dumas, and Peter Gardner. Resonant Mie scattering in infrared spectroscopy of biological materials understanding the dispersion artefact'. *The Analyst*, 134(8):1586,

2009. ISSN 0003-2654. doi: 10.1039/b904808a. URL http://xlink.rsc.org/?DOI=b904808a.

[112] Paul Bassan, Achim Kohler, Harald Martens, Joe Lee, Hugh J. Byrne, Paul Dumas, Ehsan Gazi, Michael Brown, Noel Clarke, and Peter Gardner. Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples. *The Analyst*, 135(2):268–277, 2010. ISSN 0003-2654. doi: 10.1039/B921056C. URL http://xlink.rsc.org/?DOI=B921056C.

[113] Paul Bassan, Ashwin Sachdeva, Achim Kohler, Caryn Hughes, Alex Henderson, Jonathan Boyle, Jonathan H. Shanks, Michael Brown, Noel W. Clarke, and Peter Gardner. FTIR microscopy of biological cells and tissue: data analysis using resonant Mie scattering (RMieS) EMSC algorithm. *The Analyst*, 137(6):1370, 2012. ISSN 0003-2654. doi: 10.1039/c2an16088a. URL http://xlink.rsc.org/?DOI=c2an16088a.

[114] Brian Mohlenhoff, Melissa Romeo, Max Diem, and Bayden R. Wood. Mie-Type Scattering and Non-Beer-Lambert Absorption Behavior of Human Cells in Infrared Microspectroscopy. *Biophysical Journal*, 88(5):3635–3640, 2005. ISSN 00063495. doi: 10.1529/biophysj.104.057950. URL http://linkinghub.elsevier.com/retrieve/pii/S0006349505734123.

[115] P. Walstra. Approximation formulae for the light scattering coefficient of dielectric spheres. *British Journal of Applied Physics*, 15(12):1545–1552, 1964. ISSN 05083443. doi: 10.1088/0508-3443/15/12/315.

[116] Gardner P. Bassan P, Kohler A, Martens H, Lee J, Jackson E, Lockyer N, Dumas P, Brown M, Clarke N. RMieS-EMSC correction for infrared spectra of biological cells: extension using full Mie theory and GPU computing. *Biophotonics*, 3(8):609–620, 2010.

[117] Roy M. Bremnes, Tom Dønnem, Samer Al-Saad, Khalid Al-Shibli, Sigve Andersen, Rafael Sirera, Carlos Camps, Inigo Marinez, and Lill-Tove

Busund. The Role of Tumor Stroma in Cancer Progression and Prognosis: Emphasis on Carcinoma-Associated Fibroblasts and Non-small Cell Lung Cancer. *Journal of Thoracic Oncology*, 6(1):209–217, 2011. ISSN 15560864. doi: 10.1097/JTO.0b013e3181f8a1bd. URL `http://linkinghub.elsevier.com/retrieve/pii/S1556086415319201`.

[118] P T Wong, R K Wong, T a Caputo, T a Godwin, and B Rigas. Infrared spectroscopy of exfoliated human cervical cells: evidence of extensive structural changes during carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 88(24):10988–10992, 1991. ISSN 0027-8424. doi: 10.1073/pnas.88.24.10988.

[119] Chris Holmberg, Michael Quante, Islay Steele, Jothi Dinesh Kumar, Silviya Balabanova, Cedric Duval, Matyas Czepan, Zoltan Rakonczay, Laszlo Tiszlavicz, Istvan Nemeth, Gyorgy Lazar, Zsolt Simonka, Rosalind Jenkins, Peter Hegyi, Timothy C. Wang, Graham J. Dockray, and Andrea Varro. Release of TGFβig-h3 by gastric myofibroblasts slows tumor growth and is decreased with cancer progression. *Carcinogenesis*, 33(8):1553–1562, 2012. ISSN 01433334. doi: 10.1093/carcin/bgs180.

[120] Desbordes Paul, Ruan Su, Modzelewski Romain, Vauclin Sébastien, Vera Pierre, and Gardin Isabelle. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Computerized Medical Imaging and Graphics*, 60:42–49, 2017. ISSN 18790771. doi: 10.1016/j.compmedimag.2016.12.002. URL `http://dx.doi.org/10.1016/j.compmedimag.2016.12.002`.

[121] Jannis Heil, Xandra Michaelis, Bernd Marschner, and Britta Stumpe. The power of Random Forest for the identification and quantification of technogenic substrates in urban soils on the basis of DRIFT spectra. *Environmental Pollution*, 230:574–583, 2017. ISSN 18736424. doi: 10.1016/j.envpol.2017.06.086. URL `http://dx.doi.org/10.1016/j.envpol.2017.06.086`.

[122] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015. ISSN 20010370. doi: 10.1016/j.csbj.2014.11.005. URL `http://dx.doi.org/10.1016/j.csbj.2014.11.005`.

[123] Random forest source code. URL `https://github.com/tingliu/randomforest-matlab`.